Durham University

互文性和文体分析

德龙 (Donald Sturgeon)
英国杜伦大学
计算机科学系

相关资料和PPT下载：https://dsturgeon.net/pku2023

概要

A.	正则表达式	Regular expressions
B.	互文性	Intertextuality
C.	文体分析	Stylometry

相关资料和PPT下载：https://dsturgeon.net/pku2023

互文性（Intertextuality）

1. 什么是「互文性」？
2. 动机：实际的研究案例
3. 技术：用什么样的技术手段把互文性给找出来?
4. 实践：大家一起来体验一下吧!

Durham University

相关资料和PPT下载：https://dsturgeon.net/pku2023

1. 什么是互文性?

互文性：文本间的非偶然相似性　（我今天的定义，仅供参考！）

例如

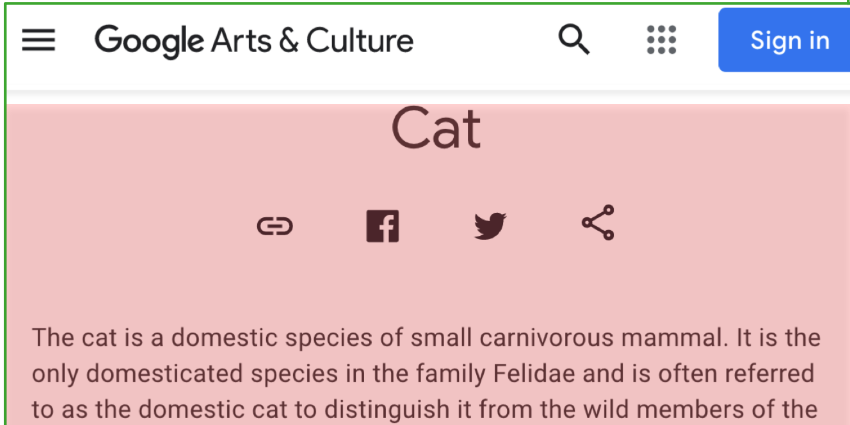
- 文本A引用了文本B的内容
- 文本A抄袭了文本B的内容
- 文本A故意让读者想起文本B的内容　=>　Allusion（暗示、典故）
- 文本A和文本B有作者潜意识留下来的相似度　=>　Stylometry（计量文体学）

Text reuse
(文本再利用)

前两种比较容易用简单数子方法找出来

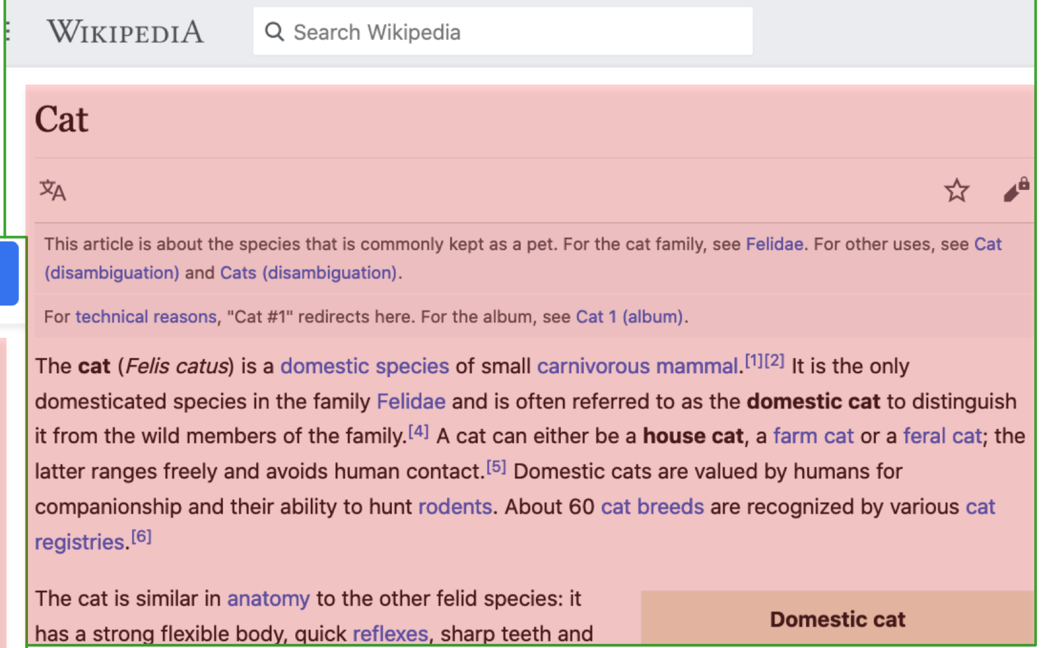
后两种也可以用电脑处理，但难度高一些

1. 什么是互文性? Global text reuse



Cat

The cat is a domestic species of small carnivorous mammal. It is the only domesticated species in the family Felidae and is often referred to as the domestic cat to distinguish it from the wild members of the family. A cat can either be a house cat, a
latter ranges freely and avoids human co
valued by humans for companionship ar
About 60 cat breeds are recognized by v
The cat is similar in anatomy to the othe



Cat

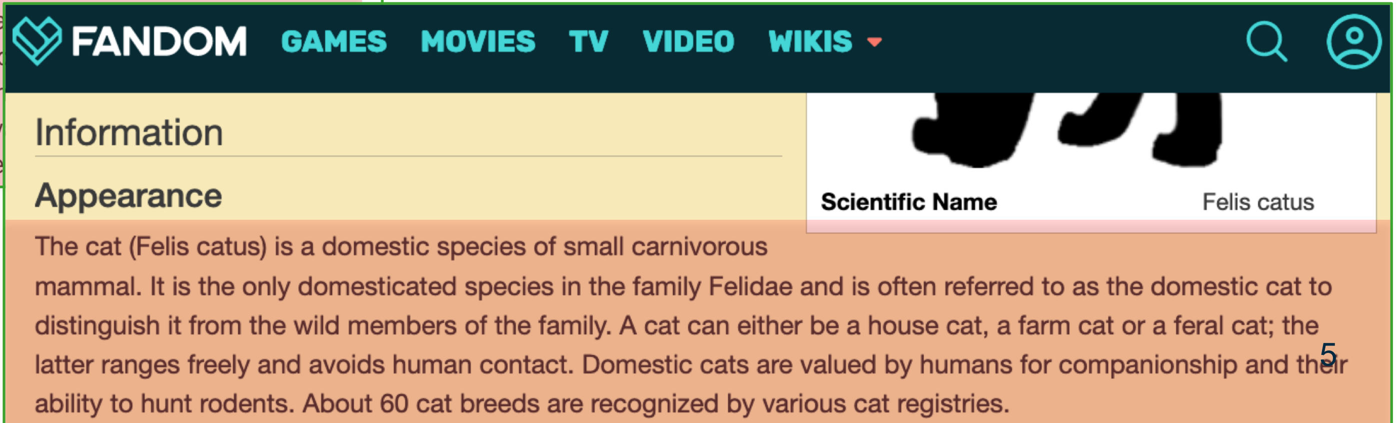
This article is about the species that is commonly kept as a pet. For the cat family, see Felidae. For other uses, see Cat (disambiguation) and Cats (disambiguation).

For technical reasons, "Cat #1" redirects here. For the album, see Cat 1 (album).

The **cat** (*Felis catus*) is a domestic species of small carnivorous mammal.[1][2] It is the only domesticated species in the family Felidae and is often referred to as the **domestic cat** to distinguish it from the wild members of the family.[4] A cat can either be a **house cat**, a farm cat or a feral cat; the latter ranges freely and avoids human contact.[5] Domestic cats are valued by humans for companionship and their ability to hunt rodents. About 60 cat breeds are recognized by various cat registries.[6]

The cat is similar in anatomy to the other felid species: it has a strong flexible body, quick reflexes, sharp teeth and

Domestic cat



Information

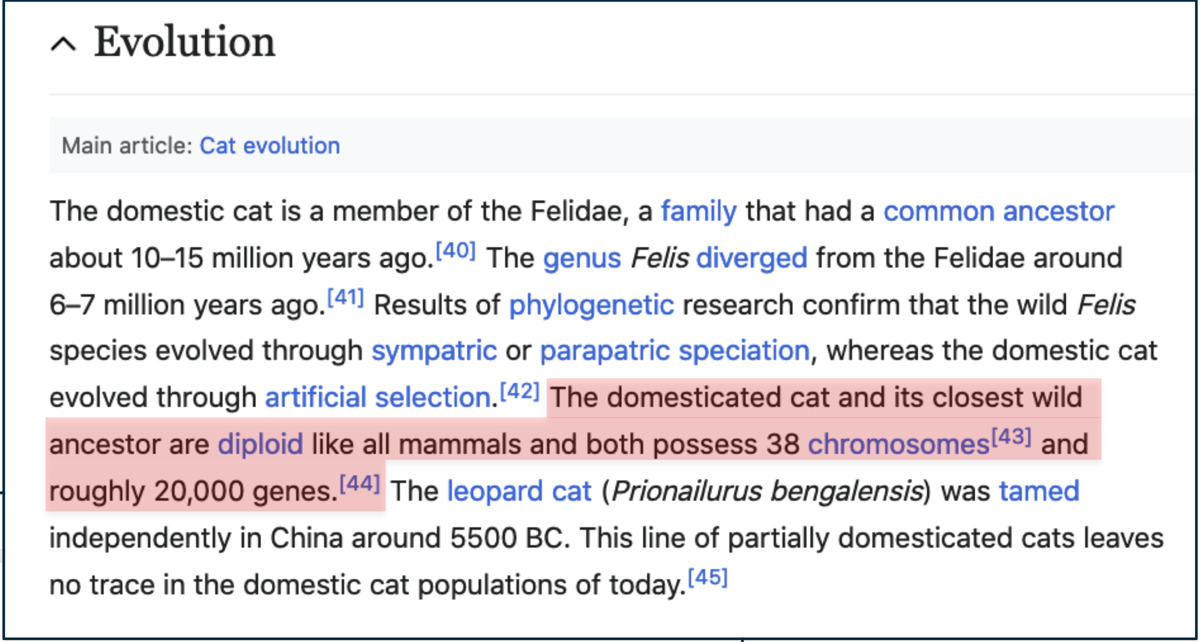
Appearance

Scientific Name Felis catus

The cat (Felis catus) is a domestic species of small carnivorous mammal. It is the only domesticated species in the family Felidae and is often referred to as the domestic cat to distinguish it from the wild members of the family. A cat can either be a house cat, a farm cat or a feral cat; the latter ranges freely and avoids human contact. Domestic cats are valued by humans for companionship and their ability to hunt rodents. About 60 cat breeds are recognized by various cat registries.

Durham University

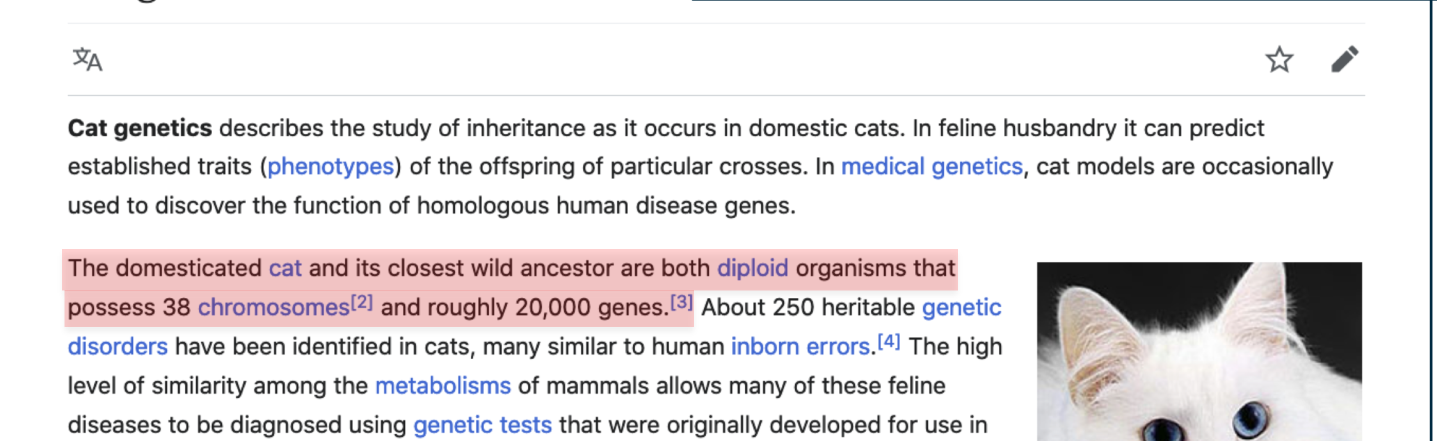
1. 什么是互文性？
Local reuse



Evolution

Main article: Cat evolution

The domestic cat is a member of the Felidae, a family that had a common ancestor about 10–15 million years ago.[40] The genus *Felis* diverged from the Felidae around 6–7 million years ago.[41] Results of phylogenetic research confirm that the wild *Felis* species evolved through sympatric or parapatric speciation, whereas the domestic cat evolved through artificial selection.[42] The domesticated cat and its closest wild ancestor are diploid like all mammals and both possess 38 chromosomes[43] and roughly 20,000 genes.[44] The leopard cat (*Prionailurus bengalensis*) was tamed independently in China around 5500 BC. This line of partially domesticated cats leaves no trace in the domestic cat populations of today.[45]



Cat genetics

Cat genetics describes the study of inheritance as it occurs in domestic cats. In feline husbandry it can predict established traits (phenotypes) of the offspring of particular crosses. In medical genetics, cat models are occasionally used to discover the function of homologous human disease genes.

The domesticated cat and its closest wild ancestor are both diploid organisms that possess 38 chromosomes[2] and roughly 20,000 genes.[3] About 250 heritable genetic disorders have been identified in cats, many similar to human inborn errors.[4] The high level of similarity among the metabolisms of mammals allows many of these feline diseases to be diagnosed using genetic tests that were originally developed for use in

1. 什么是互文性?

學而

子曰：「學而時習之，不亦說乎？有朋自遠方來，不亦樂乎？人不知而不慍，不亦君子乎？」

有子曰：「其為人也孝弟，而好犯上者，鮮矣；不好犯上，而好作亂者，未之有也。君子務本，本立而道生。孝弟也者，其為仁之本與！」

子曰：「巧言令色，鮮矣仁。」

曾子曰：「吾日三省吾身：為人謀而不忠乎？與朋友交而不信乎？傳不習乎？」

陽貨

子曰：「古者民有三疾，今也或是之亡也。古之狂也肆，今之狂也蕩；古之矜也廉，今之矜也忿戾；古之愚也直，今之愚也詐而已矣。」

子曰：「巧言令色，鮮矣仁。」

子曰：「惡紫之奪朱也，惡鄭聲之亂雅樂也，惡利口之覆邦家者。」

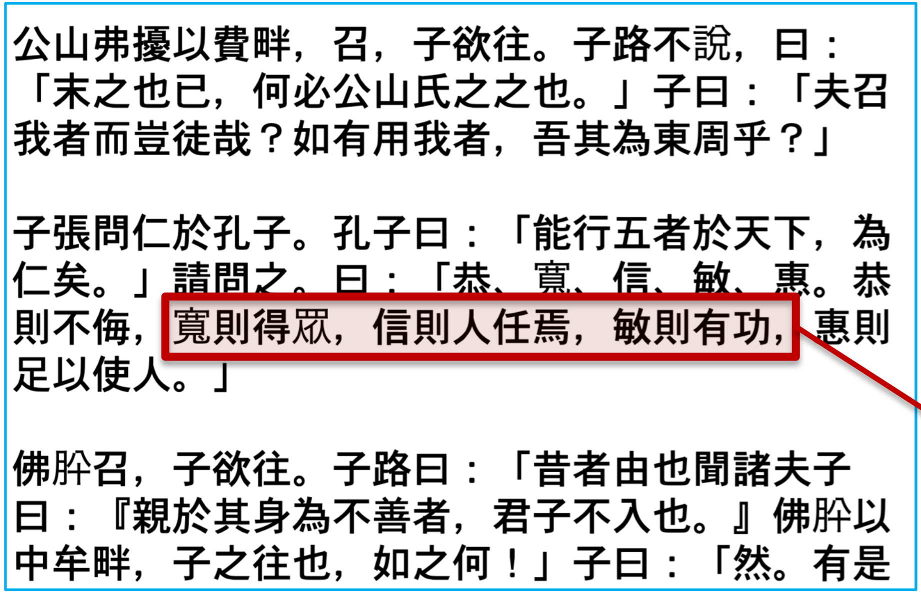
子曰：「予欲無言。」子貢曰：「子如不言，則小子何述焉？」子曰：「天何言哉？四時行焉，百物生焉，天何言哉？」

1. 什么是互文性?

恭則不侮，寬則得眾，信則人任焉，敏則有功，惠則足以使人。

所重：民、食、喪、祭。寬則得眾，信則民任焉，敏則有功，公則說。

陽貨

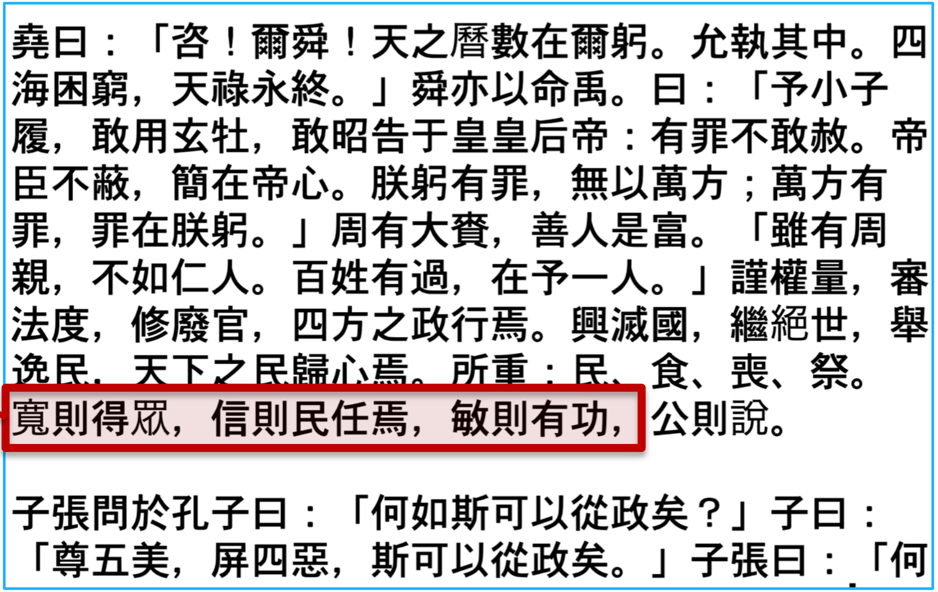


公山弗擾以費畔，召，子欲往。子路不說，曰：「末之也已，何必公山氏之之也。」子曰：「夫召我者而豈徒哉？如有用我者，吾其為東周乎？」

子張問仁於孔子。孔子曰：「能行五者於天下，為仁矣。」請問之。曰：「恭、寬、信、敏、惠。恭則不侮，寬則得眾，信則人任焉，敏則有功，惠則足以使人。」

佛肸召，子欲往。子路曰：「昔者由也聞諸夫子曰：『親於其身為不善者，君子不入也。』佛肸以中牟畔，子之往也，如之何！」子曰：「然。有是

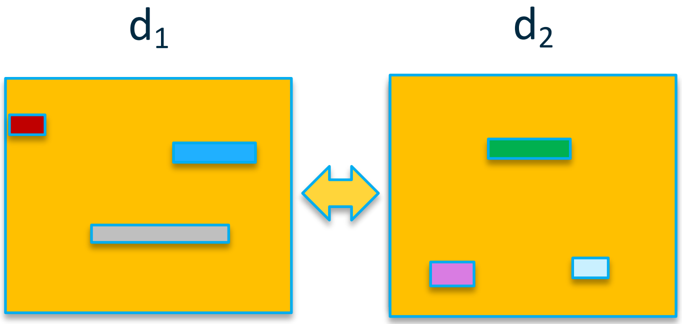
堯曰



堯曰：「咨！爾舜！天之曆數在爾躬。允執其中。四海困窮，天祿永終。」舜亦以命禹。曰：「予小子履，敢用玄牡，敢昭告于皇皇后帝：有罪不敢赦。帝臣不蔽，簡在帝心。朕躬有罪，無以萬方；萬方有罪，罪在朕躬。」周有大賚，善人是富。「雖有周親，不如仁人。百姓有過，在予一人。」謹權量，審法度，修廢官，四方之政行焉。興滅國，繼絕世，舉逸民，天下之民歸心焉。所重：民、食、喪、祭。寬則得眾，信則民任焉，敏則有功，公則說。

子張問於孔子曰：「何如斯可以從政矣？」子曰：「尊五美，屏四惡，斯可以從政矣。」子張曰：「何

Global vs local reuse

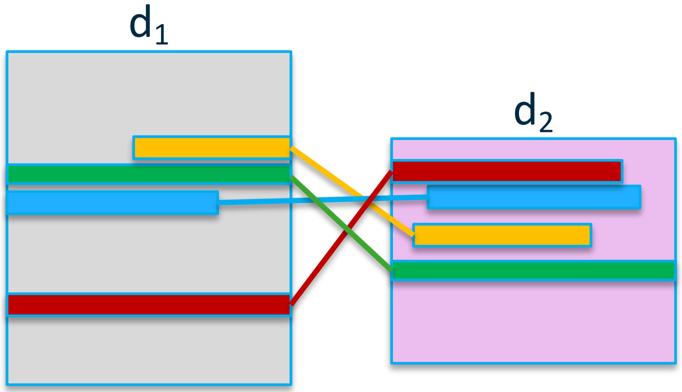


Global

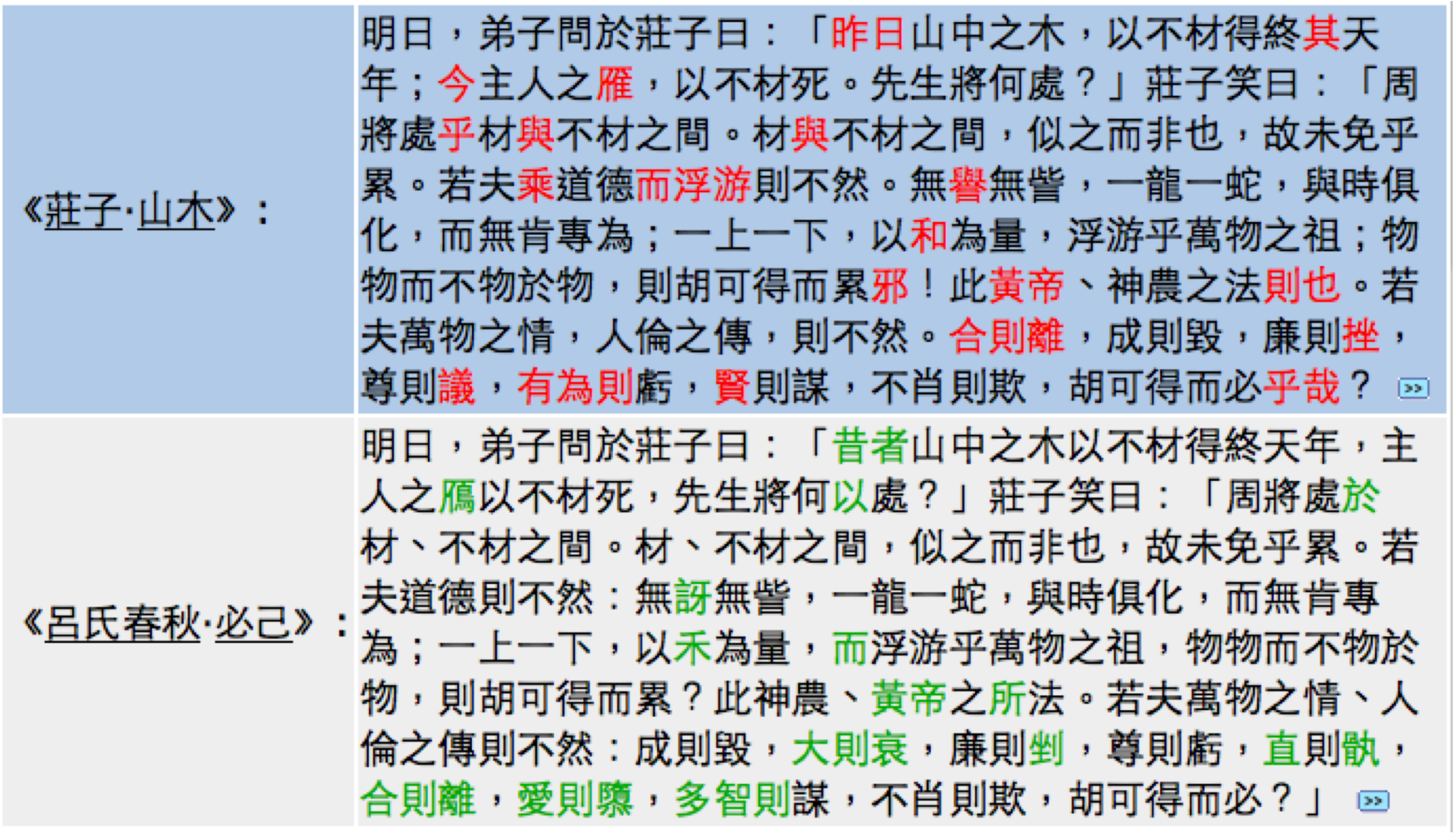
- Usually a question of similarity of *documents*
- May not require alignment of similar parts

Local

- (Possibly) isolated regions of d_1 and d_2
- Often multiple similarities between d_1 and d_2
- Similar regions may not occur in order*
- Many applications require local alignments



* If we know that they do, or if we only need the most similar pair, things get much easier



《莊子·山木》：	明日，弟子問於莊子曰：「昨日山中之木，以不材得終其天年；今主人之雁，以不材死。先生將何處？」莊子笑曰：「周將處乎材與不材之間。材與不材之間，似之而非也，故未免乎累。若夫乘道德而浮游則不然。無譽無訾，一龍一蛇，與時俱化，而無肯專為；一上一下，以和為量，浮游乎萬物之祖；物物而不物於物，則胡可得而累邪！此黃帝、神農之法則也。若夫萬物之情，人倫之傳，則不然。合則離，成則毀，廉則挫，尊則議，有為則虧，賢則謀，不肖則欺，胡可得而必乎哉？ >>
《呂氏春秋·必己》：	明日，弟子問於莊子曰：「昔者山中之木以不材得終天年，主人之鴈以不材死，先生將何以處？」莊子笑曰：「周將處於材、不材之間。材、不材之間，似之而非也，故未免乎累。若夫道德則不然：無訝無訾，一龍一蛇，與時俱化，而無肯專為；一上一下，以禾為量，而浮游乎萬物之祖，物物而不物於物，則胡可得而累？此神農、黃帝之所法。若夫萬物之情、人倫之傳則不然：成則毀，大則衰，廉則剉，尊則虧，直則骫，合則離，愛則隳，多智則謀，不肖則欺，胡可得而必？」 >>

2. 动机：为什么看互文性？

例子：先秦文献中的「相似段落」/「重见资料」

传统人文研究也十分重视互文性

用传统方法，找出互文性并不简单

电脑能不能：

1. 帮我们找的更快？
2. 帮我们找的更好？
3. 如果能找的好，能不能帮我们宏观看互文性现象在文本间的特性？

Unsupervised identification of text reuse in early Chinese literature
https://dsturgeon.net/text-reuse-chinese-literature/

Durham University

研究案例：先秦两汉传世文献的文本再利用

研究目标：

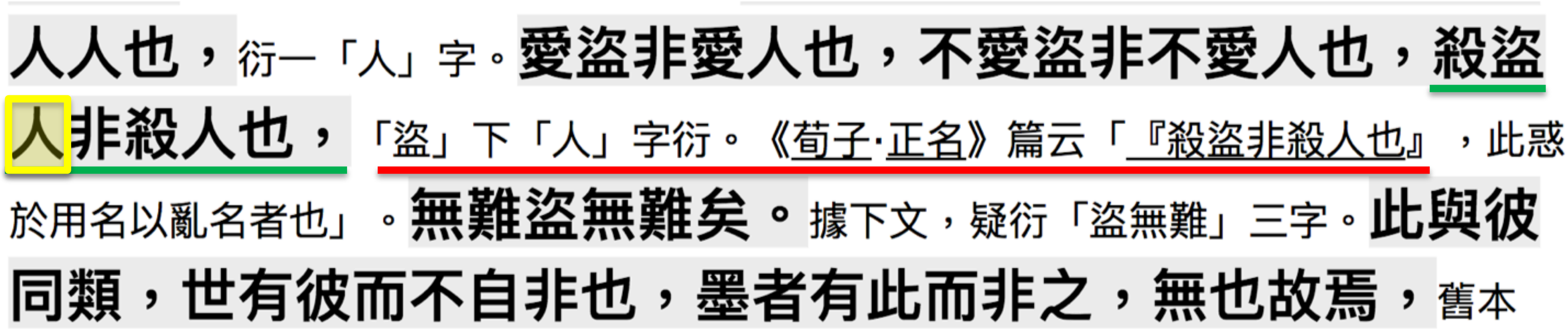
- 电脑能准确找出「相似段落」吗？
    - 怎么评估结果的正确性？
- 先秦中各种类型古籍文献之间的相似段落情况如何？
    - 哪些古籍跟哪些古籍有较多/较少的互文性？
        - 怎么给这个问题一种具体的呈现？

2. 动机：为什么看互文性？

墨子 …愛盜非愛人也；不愛盜非不愛人也；殺盜人非殺人也， …

荀子 … 「聖人不愛己」，「殺盜非殺人也」，此惑於用名以亂名者也。 …

（清）孫詒讓《墨子閒詁》：



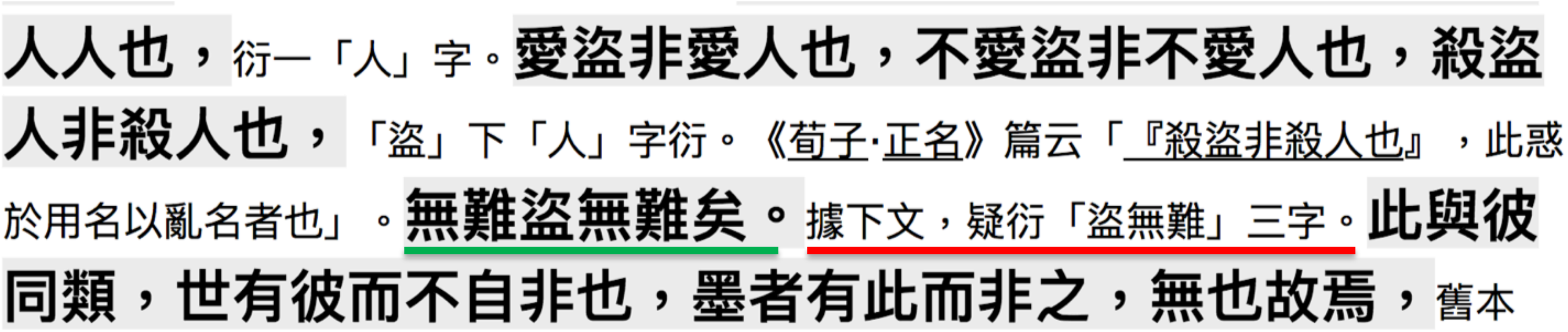
人人也，衍一「人」字。愛盜非愛人也，不愛盜非不愛人也，殺盜人非殺人也，「盜」下「人」字衍。《荀子·正名》篇云「『殺盜非殺人也』，此惑於用名以亂名者也」。無難盜無難矣。據下文，疑衍「盜無難」三字。此與彼同類，世有彼而不自非也，墨者有此而非之，無也故焉，舊本

《荀子·正名》：

「見侮不辱」，「聖人不愛己」，「殺盜非殺人也」，此惑於用名以亂名者也。驗之所為有名，而觀其孰行，則能禁之矣。「山淵平」，「情欲寡」，「芻豢不加甘，大鐘不加樂」，此惑於用實，以亂名者也。…

2. 动机：为什么看互文性？

（清）孫詒讓《墨子閒詁》：



人人也，衍一「人」字。**愛盜非愛人也，不愛盜非不愛人也，殺盜人非殺人也，**「盜」下「人」字衍。《荀子·正名》篇云「『殺盜非殺人也』，此惑於用名以亂名者也」。**無難盜無難矣。**據下文，疑衍「盜無難」三字。**此與彼同類，世有彼而不自非也，墨者有此而非之，無也故焉，**舊本

1 …人人也，愛盜非愛人也；不愛盜非不愛人也；殺盜人非殺人也，無難盜無難矣。此與彼同類，世有彼而不自非也，墨者有此而非之，無也故焉，…

2 ……有命，非命也；非執有命，非命也，無難矣。此與彼同，世有彼而不自非也，墨者有此而罪非之，無也故焉，所謂內膠外閉與心毋空乎？內膠而不解也。…

2. 动机：为什么看互文性？
例子：先秦文献中的「相似段落」/「重见资料」

问题一：我这次想找的「互文性」是什么？

- 论文中的说明：「For the purposes of this study, ‘parallel passages’ are defined as sections of text which contain significant common subsequences of characters having a high degree of similarity, and in addition appear likely to be causally related in terms of word choice by something more than the writers shared linguistic competence of the language.」

…今予發惟恭行天之罰。今日之事，…

⇕

…予非爾田野葆士之欲也，予共行天之罰也。左不共于左， …

…必謹所堪者，此之謂也。…

⇕

…《書》云:「凡人自得罪。」此之謂也。…

2. 动机：为什么看互文性？
例子：先秦文献中的「相似段落」/「重见资料」

问题二：怎么让电脑去找出符合这种定义的相似段落？

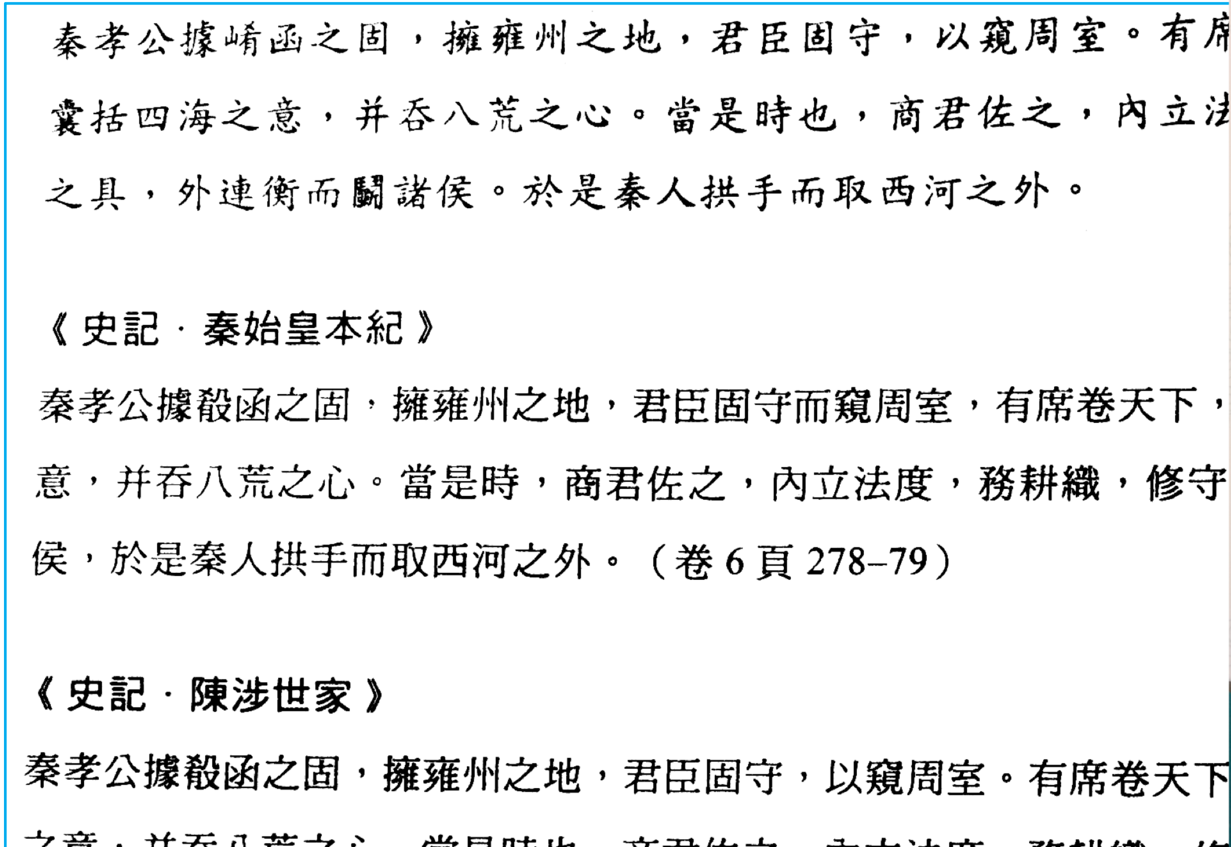
- 这个等会儿再说（详细说明可以参考论文）

问题三：怎么证明电脑多么准确找到了这种互文性？

- 看电脑所找出来的例子是不是好例子…这样够吗？
  - 不够：还要看有没有漏掉了很多例子
- 如果有现成的比较完整的数据或参考书，可以抽出一部分系统地比较
  => 所谓的「Test set」

2. 动机：为什么看互文性？

例子：先秦文献中的「相似段落」/「重见资料」



秦孝公據崤函之固，擁雍州之地，君臣固守，以窺周室。有席

囊括四海之意，并吞八荒之心。當是時也，商君佐之，內立法

之具，外連衡而鬬諸侯。於是秦人拱手而取西河之外。

《史記·秦始皇本紀》

秦孝公據殽函之固，擁雍州之地，君臣固守而窺周室，有席卷天下，

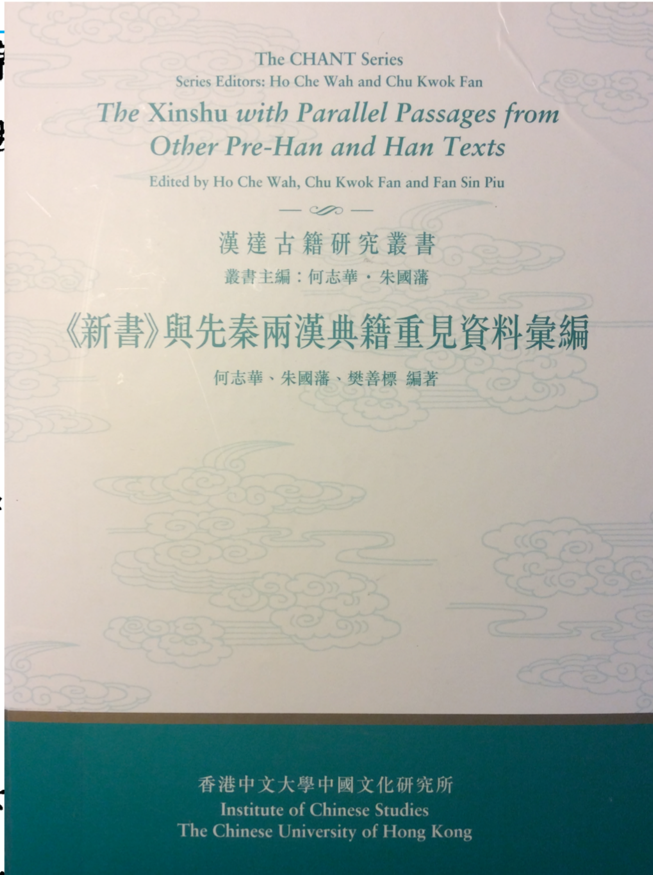
意，并吞八荒之心。當是時，商君佐之，內立法度，務耕織，修守

侯，於是秦人拱手而取西河之外。（卷 6 頁 278–79）

《史記·陳涉世家》

秦孝公據殽函之固，擁雍州之地，君臣固守，以窺周室。有席卷天下

之意，并吞八荒之心。當是時也，商君佐之，內立法度，務耕織，修



The CHANT Series
Series Editors: Ho Che Wah and Chu Kwok Fan

The Xinshu with Parallel Passages from Other Pre-Han and Han Texts

Edited by Ho Che Wah, Chu Kwok Fan and Fan Sin Piu

漢達古籍研究叢書
叢書主編：何志華、朱國藩

《新書》與先秦兩漢典籍重見資料彙編

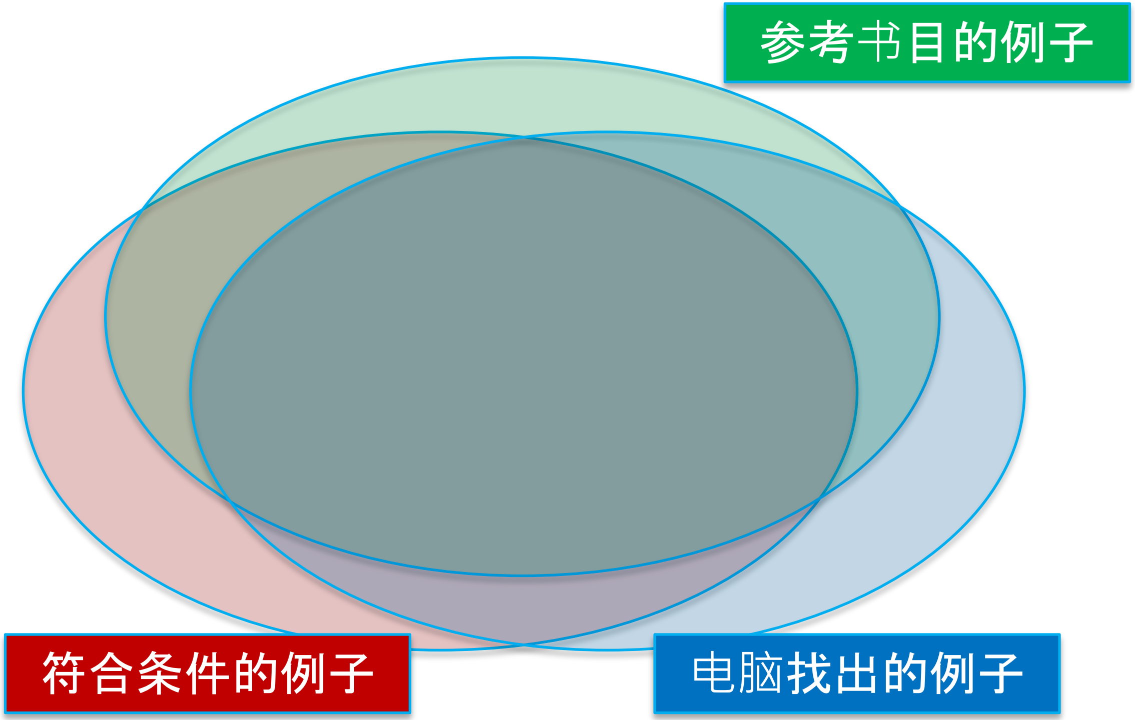
何志華、朱國藩、樊善標 編著

香港中文大學中國文化研究所
Institute of Chinese Studies
The Chinese University of Hong Kong

2. 动机：为什么看互文性？

例子：先秦文献中的「相似段落」/「重见资料」

怎么证明电脑多么准确找到了这中互文性？



我到底想要什么？！
或者说：电脑算出来的结果是对的吗？

- 问题一：我希望电脑可以找出来哪一些相似的例子？
- 问题二：电脑实际上是否找到了我想要的那些例子？

1. 有一些我想要的，电脑没找到！ 「false negative」
2. 有一些我不想要的，电脑反而说有！ 「false positive」

1. Precision（精准率）：电脑说的那些例子，百分之多少是对的？
2. Recall（召回率）：实际上存在的那些例子，电脑找到了百分之多少？

我到底想要什么？！
或者说：电脑算出来的结果是对的吗？

與香港中文大學ICS《荀子》首篇的比較

	精確率	召回率
ICS (寬鬆)	100%	54%
CTP (寬鬆)	100%	94%

我到底想要什么？！
或者说：电脑算出来的结果是对的吗？

参考书目落掉了的例子：

順風而呼，聲非加疾也，	《荀子》
順風而呼，聲不加疾也；	《呂氏春秋》
比如順風而呼，聲非加疾，	《史記》

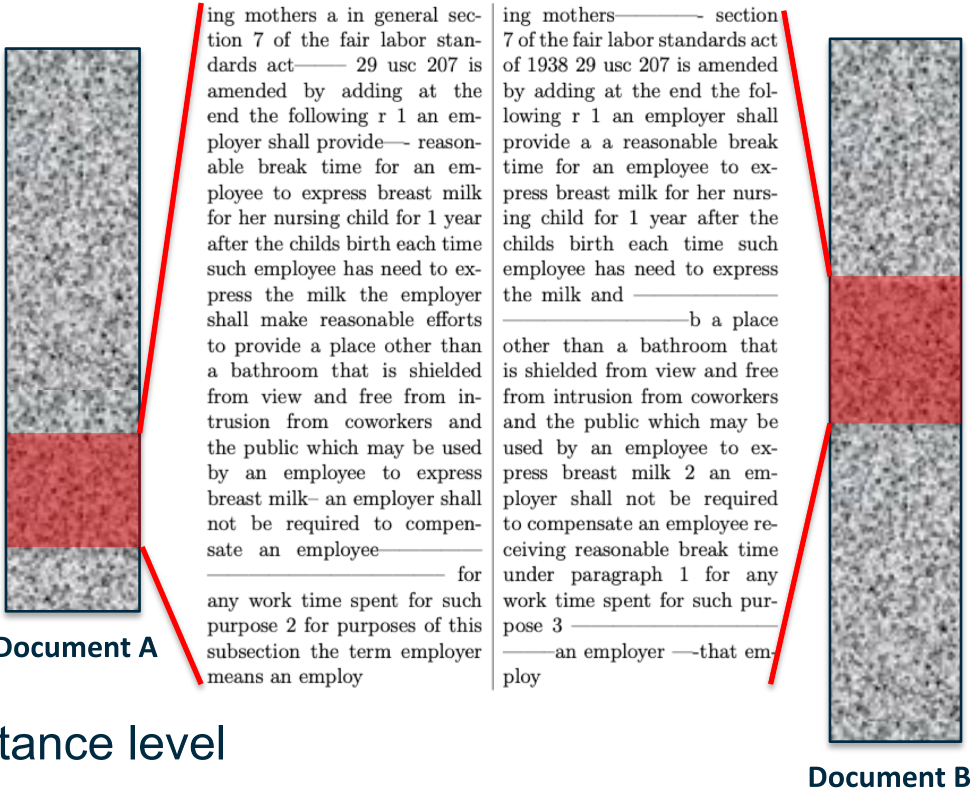
青取之於藍，而青於藍；	《荀子》
青采出於藍，而質青於藍	《史記》

古之學者為己，今之學者為人	《荀子》
古之學者為己，今之學者為人	《論語》

互文性的可视化

Highly specific to corpus & reuse

- Is reuse (mostly?) sequential?
- What are the aligned units?
- At what scales will it be done?
  - Between works
  - Between parts of works
  - Between lines of text, …
- Alignments expensive
  - Highly interpretable on an instance level



ing mothers a in general section 7 of the fair labor standards act——— 29 usc 207 is amended by adding at the end the following r 1 an employer shall provide—- reasonable break time for an employee to express breast milk for her nursing child for 1 year after the childs birth each time such employee has need to express the milk the employer shall make reasonable efforts to provide a place other than a bathroom that is shielded from view and free from intrusion from coworkers and the public which may be used by an employee to express breast milk– an employer shall not be required to compensate an employee——————— for any work time spent for such purpose 2 for purposes of this subsection the term employer means an employ

ing mothers———- section 7 of the fair labor standards act of 1938 29 usc 207 is amended by adding at the end the following r 1 an employer shall provide a a reasonable break time for an employee to express breast milk for her nursing child for 1 year after the childs birth each time such employee has need to express the milk and ——————b a place other than a bathroom that is shielded from view and free from intrusion from coworkers and the public which may be used by an employee to express breast milk 2 an employer shall not be required to compensate an employee receiving reasonable break time under paragraph 1 for any work time spent for such purpose 3 ——————an employer —that employ

Durham University

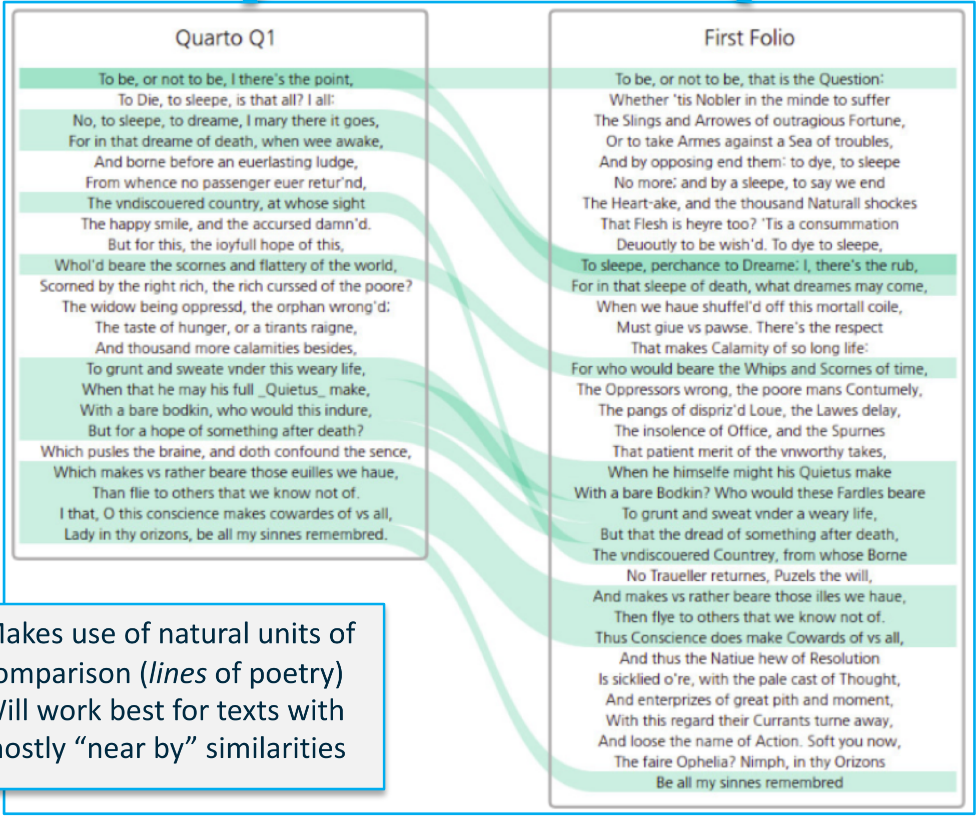
https://www.jstor.org/stable/24582958

文性的可视化

Small-scale alignments

- Highlighting of text
- Connecting similar text
  - Including reordering
- Hard to directly scale
  - Breaks down when text is outside visual field

Two different early editions of the *same* work by Shakespeare



Quarto Q1

To be, or not to be, I there's the point,
To Die, to sleepe, is that all? I all:
No, to sleepe, to dreame, I mary there it goes,
For in that dreame of death, when wee awake,
And borne before an euerlasting Iudge,
From whence no passenger euer retur'nd,
The vndiscouered country, at whose sight
The happy smile, and the accursed damn'd.
But for this, the ioyfull hope of this,
Whol'd beare the scornes and flattery of the world,
Scorned by the right rich, the rich curssed of the poore?
The widow being oppressd, the orphan wrong'd:
The taste of hunger, or a tirants raigne,
And thousand more calamities besides,
To grunt and sweate vnder this weary life,
When that he may his full _Quietus_ make,
With a bare bodkin, who would this indure,
But for a hope of something after death?
Which pusles the braine, and doth confound the sence,
Which makes vs rather beare those euilles we haue,
Than flie to others that we know not of.
I that, O this conscience makes cowardes of vs all,
Lady in thy orizons, be all my sinnes remembred.

First Folio

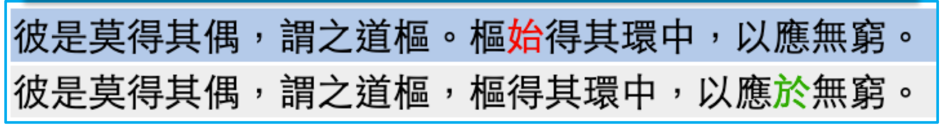
To be, or not to be, that is the Question:
Whether 'tis Nobler in the minde to suffer
The Slings and Arrowes of outragious Fortune,
Or to take Armes against a Sea of troubles,
And by opposing end them: to dye, to sleepe
No more; and by a sleepe, to say we end
The Heart-ake, and the thousand Naturall shockes
That Flesh is heyre too? 'Tis a consummation
Deuoutly to be wish'd. To dye to sleepe,
To sleepe, perchance to Dreame; I, there's the rub,
For in that sleepe of death, what dreames may come,
When we haue shuffel'd off this mortall coile,
Must giue vs pawse. There's the respect
That makes Calamity of so long life:
For who would beare the Whips and Scornes of time,
The Oppressors wrong, the poore mans Contumely,
The pangs of dispriz'd Loue, the Lawes delay,
The insolence of Office, and the Spurnes
That patient merit of the vnworthy takes,
When he himselfe might his Quietus make
With a bare Bodkin? Who would these Fardles beare
To grunt and sweat vnder a weary life,
But that the dread of something after death,
The vndiscouered Countrey, from whose Borne
No Traueller returnes, Puzels the will,
And makes vs rather beare those illes we haue,
Then flye to others that we know not of.
Thus Conscience does make Cowards of vs all,
And thus the Natiue hew of Resolution
Is sicklied o're, with the pale cast of Thought,
And enterprizes of great pith and moment,
With this regard their Currants turne away,
And loose the name of Action. Soft you now,
The faire Ophelia? Nimph, in thy Orizons
Be all my sinnes remembred

- Makes use of natural units of comparison (*lines* of poetry)
- Will work best for texts with mostly "near by" similarities

A Survey of Text Alignment Visualization ctext.org

文性的可视化

Colorization of differences within closely aligned segments of text

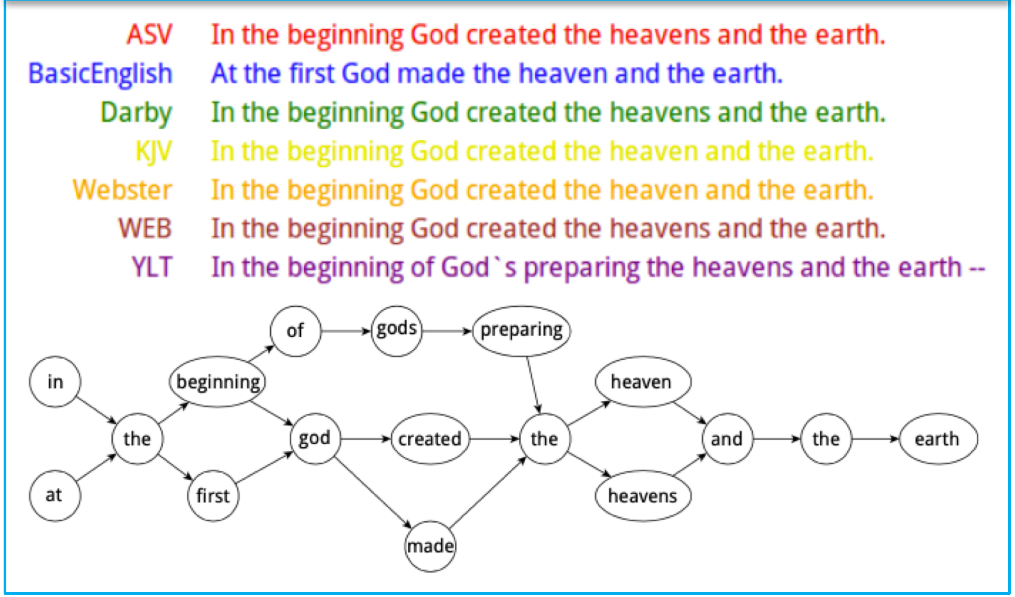


彼是莫得其偶，謂之道樞。樞始得其環中，以應無窮。

彼是莫得其偶，謂之道樞，樞得其環中，以應於無窮。

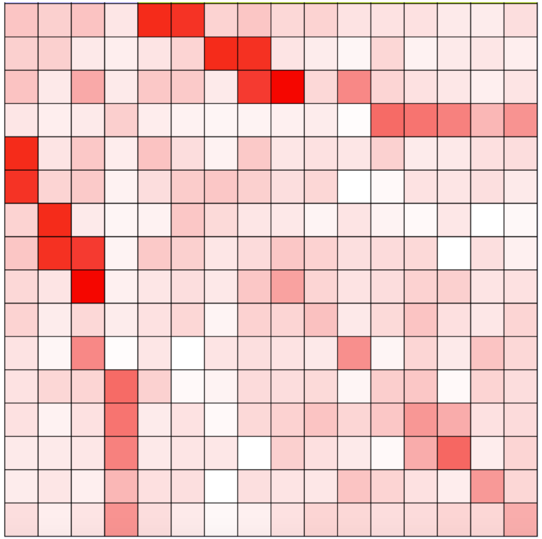
Even *smaller* scales can be useful…

Multiple (closely aligned and highly similar) alignments as a term graph

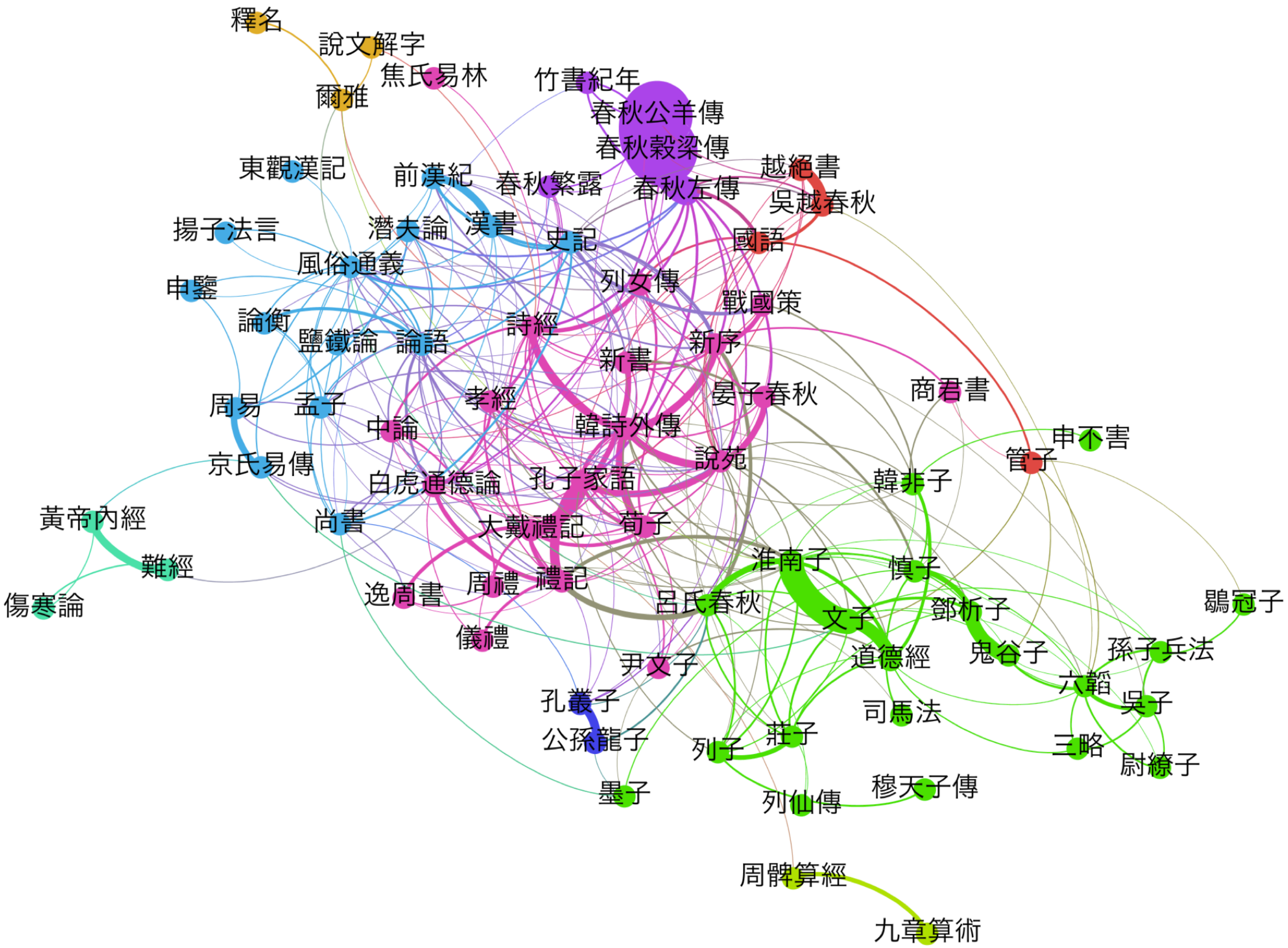


ASV	In the beginning God created the heavens and the earth.
BasicEnglish	At the first God made the heaven and the earth.
Darby	In the beginning God created the heavens and the earth.
KJV	In the beginning God created the heaven and the earth.
Webster	In the beginning God created the heaven and the earth.
WEB	In the beginning God created the heavens and the earth.
YLT	In the beginning of God`s preparing the heavens and the earth --

… and so can *larger* ones



Visualizations for text reuse



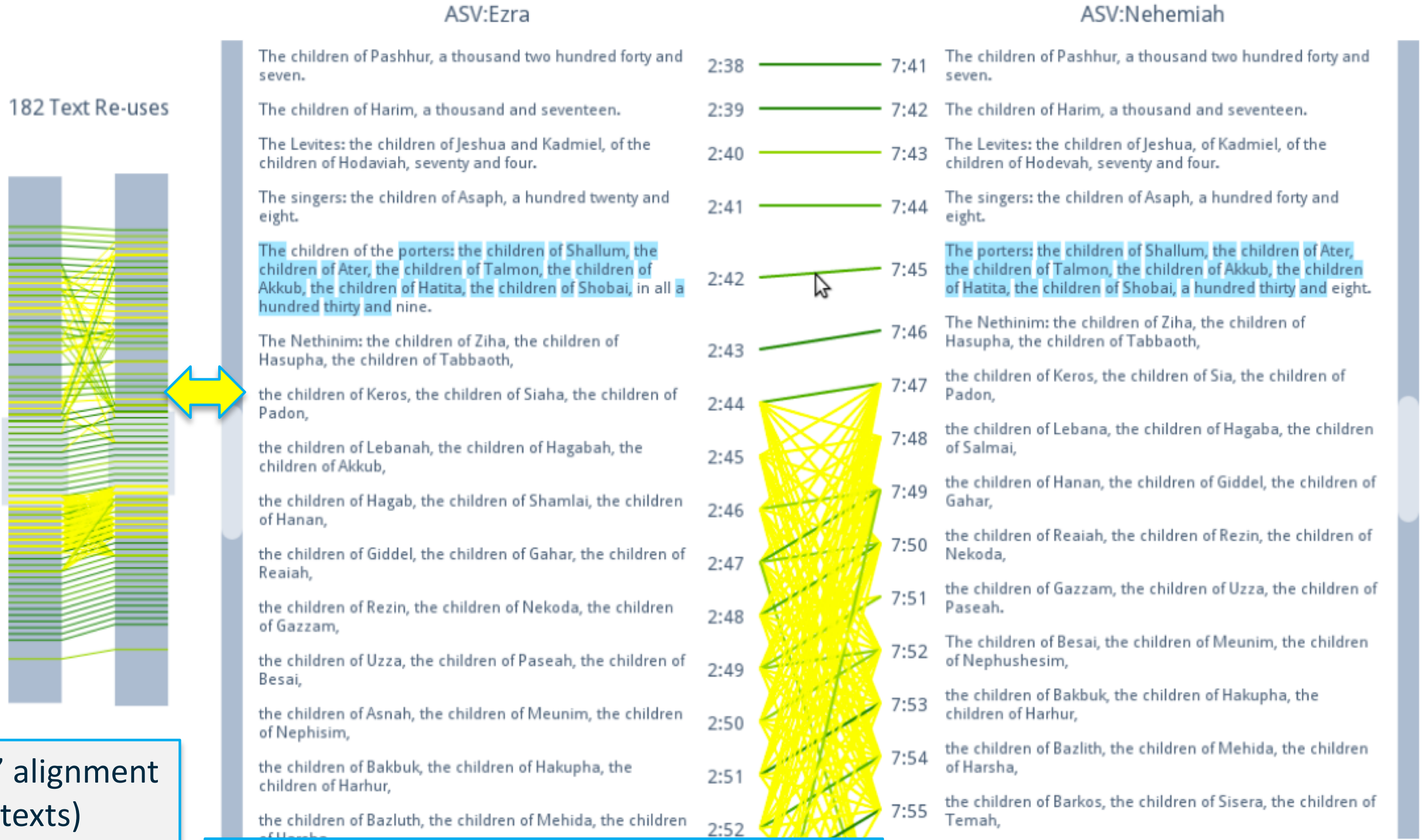
Nodes represent individual literary works; edges *summarize* reuse instances between pairs of work (e.g. 10~1000 instances per edge). Corpus size: ~5 million words

Unsupervised Identification of Text Reuse in Early Chinese Literature

Visualization of text reuse – interaction

Interactive interfaces allow navigating complex information at multiple user-selectable scales

Google



ASV:Ezra

2:38 The children of Pashhur, a thousand two hundred forty and seven.

2:39 The children of Harim, a thousand and seventeen.

2:40 The Levites: the children of Jeshua and Kadmiel, of the children of Hodaviah, seventy and four.

2:41 The singers: the children of Asaph, a hundred twenty and eight.

2:42 The children of the porters: the children of Shallum, the children of Ater, the children of Talmon, the children of Akkub, the children of Hatita, the children of Shobai, in all a hundred thirty and nine.

2:43 The Nethinim: the children of Ziha, the children of Hasupha, the children of Tabbaoth,

2:44 the children of Keros, the children of Siaha, the children of Padon,

2:45 the children of Lebanah, the children of Hagabah, the children of Akkub,

2:46 the children of Hagab, the children of Shamlai, the children of Hanan,

2:47 the children of Giddel, the children of Gahar, the children of Reaiah,

2:48 the children of Rezin, the children of Nekoda, the children of Gazzam,

2:49 the children of Uzza, the children of Paseah, the children of Besai,

2:50 the children of Asnah, the children of Meunim, the children of Nephisim,

2:51 the children of Bakbuk, the children of Hakupha, the children of Harhur,

2:52 the children of Bazluth, the children of Mehida, the children of Harsha,

ASV:Nehemiah

7:41 The children of Pashhur, a thousand two hundred forty and seven.

7:42 The children of Harim, a thousand and seventeen.

7:43 The Levites: the children of Jeshua, of Kadmiel, of the children of Hodevah, seventy and four.

7:44 The singers: the children of Asaph, a hundred forty and eight.

7:45 The porters: the children of Shallum, the children of Ater, the children of Talmon, the children of Akkub, the children of Hatita, the children of Shobai, a hundred thirty and eight.

7:46 The Nethinim: the children of Ziha, the children of Hasupha, the children of Tabbaoth,

7:47 the children of Keros, the children of Sia, the children of Padon,

7:48 the children of Lebana, the children of Hagaba, the children of Salmai,

7:49 the children of Hanan, the children of Giddel, the children of Gahar,

7:50 the children of Reaiah, the children of Rezin, the children of Nekoda,

7:51 the children of Gazzam, the children of Uzza, the children of Paseah.

7:52 The children of Besai, the children of Meunim, the children of Nephushesim,

7:53 the children of Bakbuk, the children of Hakupha, the children of Harhur,

7:54 the children of Bazlith, the children of Mehida, the children of Harsha,

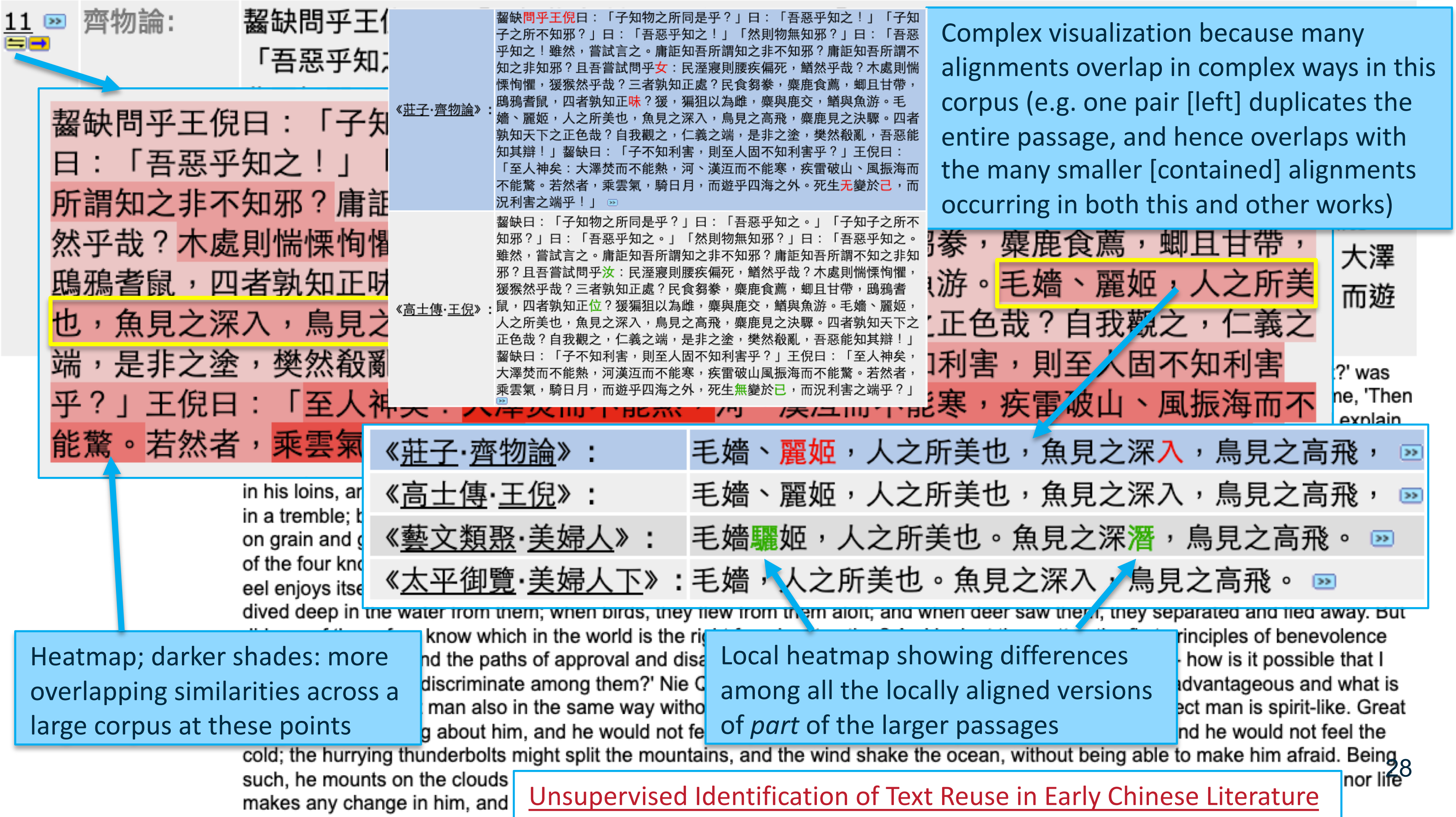
7:55 the children of Barkos, the children of Sisera, the children of Temah,

“Big picture” alignment summary (2 texts)

Textual details with fine-grained alignments

Durham University

Designing Close and Distant Reading Visualizations for Text Re-use



11 齊物論：

Complex visualization because many alignments overlap in complex ways in this corpus (e.g. one pair [left] duplicates the entire passage, and hence overlaps with the many smaller [contained] alignments occurring in both this and other works)

《莊子·齊物論》：	齧缺問乎王倪曰：「子知物之所同是乎？」曰：「吾惡乎知之！」「子知子之所不知邪？」曰：「吾惡乎知之！」「然則物無知邪？」曰：「吾惡乎知之！雖然，嘗試言之。庸詎知吾所謂知之非不知邪？庸詎知吾所謂不知之非知邪？且吾嘗試問乎女：民溼寢則腰疾偏死，鰌然乎哉？木處則惴慄恂懼，猨猴然乎哉？三者孰知正處？民食芻豢，麋鹿食薦，蝍且甘帶，鴟鴉耆鼠，四者孰知正味？猨，猵狙以為雌，麋與鹿交，鰌與魚游。毛嫱、麗姬，人之所美也，魚見之深入，鳥見之高飛，麋鹿見之決驟。四者孰知天下之正色哉？自我觀之，仁義之端，是非之塗，樊然殽亂，吾惡能知其辯！」齧缺曰：「子不知利害，則至人固不知利害乎？」王倪曰：「至人神矣：大澤焚而不能熱，河、漢沍而不能寒，疾雷破山、風振海而不能驚。若然者，乘雲氣，騎日月，而遊乎四海之外。死生无變於己，而況利害之端乎！」
《高士傳·王倪》：	齧缺曰：「子知物之所同是乎？」曰：「吾惡乎知之。」「子知子之所不知邪？」曰：「吾惡乎知之。」「然則物無知邪？」曰：「吾惡乎知之。雖然，嘗試言之。庸詎知吾所謂知之非不知邪？庸詎知吾所謂不知之非知邪？且吾嘗試問乎汝：民溼寢則腰疾偏死，鰌然乎哉？木處則惴慄恂懼，猨猴然乎哉？三者孰知正處？民食芻豢，麋鹿食薦，蝍且甘帶，鴟鴉耆鼠，四者孰知正位？猨猵狙以為雌，麋與鹿交，鰌與魚游。毛嫱、麗姬，人之所美也，魚見之深入，鳥見之高飛，麋鹿見之決驟。四者孰知天下之正色哉？自我觀之，仁義之端，是非之塗，樊然殽亂，吾惡能知其辯！」齧缺曰：「子不知利害，則至人固不知利害乎？」王倪曰：「至人神矣，大澤焚而不能熱，河漢沍而不能寒，疾雷破山風振海而不能驚。若然者，乘雲氣，騎日月，而遊乎四海之外，死生無變於已，而況利害之端乎？」

《莊子·齊物論》：	毛嫱、麗姬，人之所美也，魚見之深入，鳥見之高飛，
《高士傳·王倪》：	毛嫱、麗姬，人之所美也，魚見之深入，鳥見之高飛，
《藝文類聚·美婦人》：	毛嫱驪姬，人之所美也。魚見之深潛，鳥見之高飛。
《太平御覽·美婦人下》：	毛嫱，人之所美也。魚見之深入，鳥見之高飛。

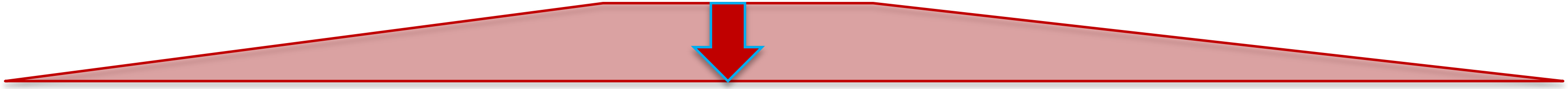
Heatmap; darker shades: more overlapping similarities across a large corpus at these points

Local heatmap showing differences among all the locally aligned versions of *part* of the larger passages

Unsupervised Identification of Text Reuse in Early Chinese Literature

Other types of reuse: literary allusion

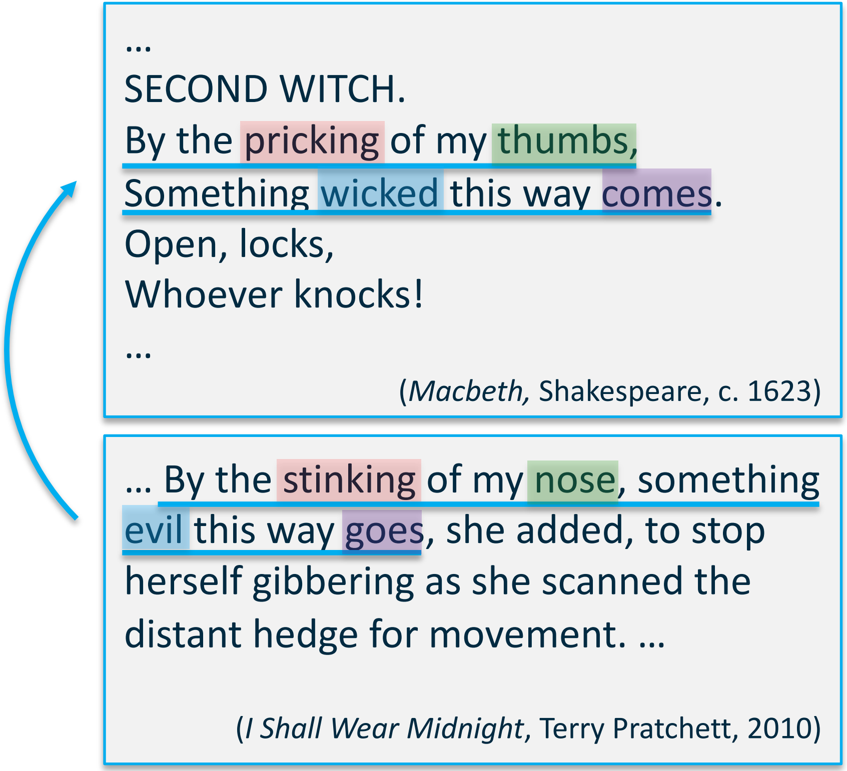
《論衡》：是故楊子哭岐道，墨子哭練絲也，蓋傷離本，不可復變也。



《墨子》：子墨子言見染絲者而歎曰：「染於蒼則蒼，染於黃則黃。所入者變，其色亦變。五入必而已，則為五色矣。故染不可不慎也。」非獨染絲然也，國亦有染。…

- 历史时间上的前后顺序
- 所指的文本具体的界限不一定完全清楚
- 有时候可能有多数可能的文本来源

Other types of reuse: literary allusion



…
SECOND WITCH.
By the pricking of my thumbs,
Something wicked this way comes.
Open, locks,
Whoever knocks!
…

(*Macbeth,* Shakespeare, c. 1623)

… By the stinking of my nose, something evil this way goes, she added, to stop herself gibbering as she scanned the distant hedge for movement. …

(*I Shall Wear Midnight*, Terry Pratchett, 2010)

Much harder to identify without false positives

Some constraints

- Directional: later => earlier (?earliest)
- Strongly biased (allusion to *famous* works)

Partly supervised approach, multiple features

- Alignment + more sophisticated scoring
- Relationships between *substituted* tokens
  - Syntactic similarities (same POS)
  - Semantic similarities (e.g. “comes”, “goes”)
- Sentence structure similarity
  - Locally parallel structures
  - Similarity of parse trees

“Shakespeare in the Vectorian Age”: detection of Shakespeare quotes

3. 技术：分词Tokenization

S_1 = 保持共产党员先进性教育活动

T_1 = 保持 共产党员 先进性 教育 活动

T_2 = 保持 共产党员 先进 性教育 活动

- Both syntactically valid tokenizations
- Different meanings
- No common rules across all languages
- Tokenization for e.g. Chinese is a non-trivial NLP task

Language specific task!

- Not trivial for all languages
  - Not necessarily as trivial as it seems even for *alphabetic* languages
  - E.g. do we really want “New York” as two tokens “New” and “York”?

3. 技术：用什么样的技术手段把互文性给找出来？

文本向量 Document vectors

- 简单、快速
- 比较适合分析global互文性
- 只看出现的词汇而忽视词出现的顺序：所谓的「Bag of words」词袋模型

N-grams

- 适用于global和local互文性
- 适合分析有连续性的互文性（如：引用、抄袭等）

其它的或自定的方法

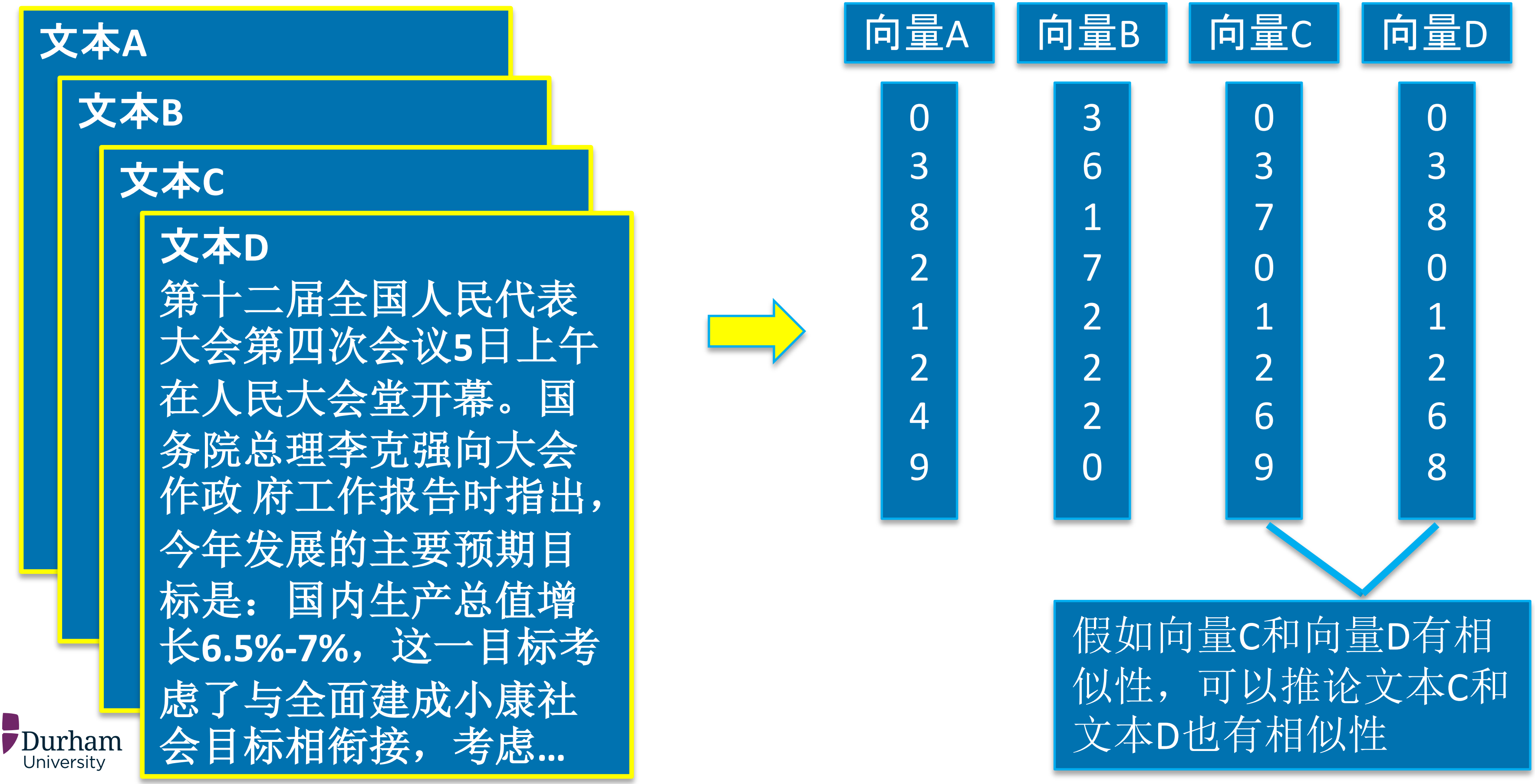
- 可以针对所想处理的互文性制定自己的标准
- 可用precision、recall等来验证、比较不同方法的有效性

3. 技术：向量 Vectors

什么是「向量」？

- 经常在自然语言处理中，我们会用一个「向量」表示一个词、文本等
- 只不过是一系列的数字而已！
- 用数字作为某一种语言表达的代替品
  - 为什么？数字是可以直接计算的；如果我们数字和它所代表的语言对象（词或文本）有某种固定关系，我们可以透过向量的计算或比较来了解语言对象之间的关系
- 每次说到向量，我们都会先定义要用多少数字（维度）
  - 一维 = 一个数字； 二维 = 两个数字； 三百位 = 三百个数字 …
- 例如：「北大」用 [10, 30, 0, 5] （四维向量）表示

3. 技术：文本向量 Document vectors



文本A

文本B

文本C

文本D

第十二届全国人民代表大会第四次会议5日上午在人民大会堂开幕。国务院总理李克强向大会作政 府工作报告时指出，今年发展的主要预期目标是：国内生产总值增长6.5%-7%，这一目标考虑了与全面建成小康社会目标相衔接，考虑…

向量A	向量B	向量C	向量D
0	3	0	0
3	6	3	3
8	1	7	8
2	7	0	0
1	2	1	1
2	2	2	2
4	2	6	6
9	0	9	8

假如向量C和向量D有相似性，可以推论文本C和文本D也有相似性

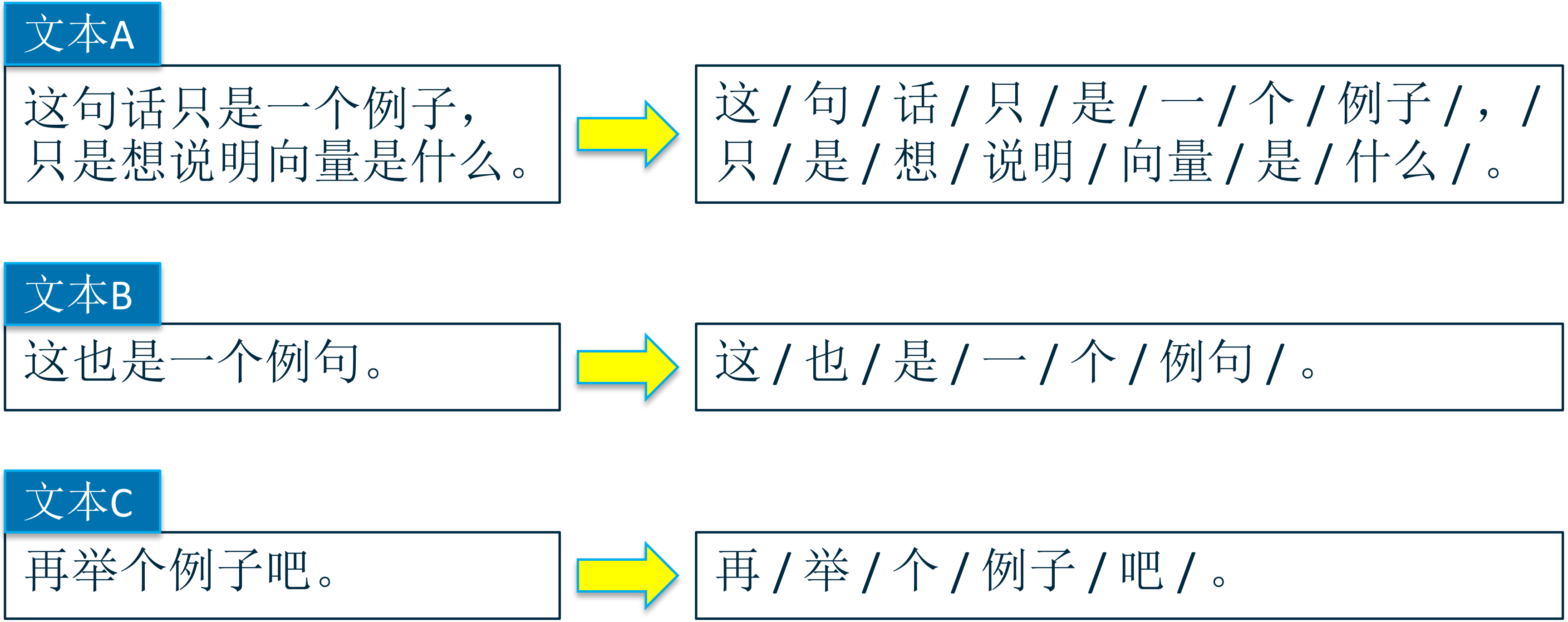
3. 技术：文本向量 Document vectors

文本向量怎么做？

1. 先制定一个「词汇」vocabulary
   - 方法一：看所有文本中出现的所有的词
   - 方法二：自定自己想分析的词汇
2. 在每一个文本里头：
   - 针对词汇中每一个项目（token，通常是词语）：
     - 写下之个token在这个文本中出现的总数

这种方法得出的向量叫做「Term frequency vectors」（词频向量）

3. 技术：文本向量实例



文本A

这句话只是一个例子，
只是想说明向量是什么。

这 / 句 / 话 / 只 / 是 / 一 / 个 / 例子 / ， /
只 / 是 / 想 / 说明 / 向量 / 是 / 什么 / 。

文本B

这也是一个例句。

这 / 也 / 是 / 一 / 个 / 例句 / 。

文本C

再举个例子吧。

再 / 举 / 个 / 例子 / 吧 / 。

3. 技术：文本向量实例

文本A

这 / 句 / 话 / 只 / 是 / 一 / 个 / 例子 / ，
/ 只 / 是 / 想 / 说明 / 向量 / 是 / 什么 / 。

文本B

这 / 也 / 是 / 一 / 个 / 例句 / 。

文本C

再 / 举 / 个 / 例子 / 吧 / 。

词汇	向量A	向量B	向量C
这	1	1	0
句	1	0	0
话	1	0	0
只	2	0	0
是	2	1	0
一	1	1	0
个	1	1	1
例子	1	0	1
，	1	0	0
想	1	0	0
说明	1	0	0
向量	1	0	0
什么	1	0	0
。	1	1	1
也	0	1	0
例句	0	1	0
举	0	0	0
吧	0	0	1

3. 技术：文本向量的相似度计算

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum\limits_{i=1}^{n} A_i B_i}{\sqrt{\sum\limits_{i=1}^{n} A_i^2}\sqrt{\sum\limits_{i=1}^{n} B_i^2}}$$

常用的向量相似度计算方法：cosine similarity

词汇	向量A	向量B	向量C
这	1	1	0
句	1	0	0
话	1	0	0
只	1	0	0
是	1	1	0
一	2	1	1
个	1	1	1
例子	1	1	0
另	0	0	1
话题	0	0	1

文本A

这 / 一 / 句 / 话 / 只 / 是 / 一 / 个 / 例子

文本B

这 / 是 / 一 / 个 / 例子

文本C

另 / 一 / 个 / 话题

3. 技术：文本向量的相似度计算

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}$$

常用的向量相似度计算方法：cosine similarity

词汇	向量A		向量B		
这	1	x	1	=	1
句	1	x	0	=	0
话	1	x	0	=	0
只	1	x	0	=	0
是	1	x	1	=	1
一	2	x	1	=	2
个	1	x	1	=	1
例子	1	x	1	=	1
另	0	x	0	=	0
话题	0	x	0	=	0
					6

文本A

这 / 一 / 句 / 话 / 只 /
是 / 一 / 个 / 例子

文本B

这 / 是 / 一 / 个 / 例子

文本C

另 / 一 / 个 / 话题

3. 技术：文本向量的相似度计算

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|} = \frac{\sum\limits_{i=1}^{n} A_i B_i}{\sqrt{\sum\limits_{i=1}^{n} A_i^2} \sqrt{\sum\limits_{i=1}^{n} B_i^2}}$$

常用的向量相似度计算方法：cosine similarity

词汇	向量A			向量B		
这	1^2	=	1	1^2	=	1
句	1^2	=	1	0^2	=	0
话	1^2	=	1	0^2	=	0
只	1^2	=	1	0^2	=	0
是	1^2	=	1	1^2	=	1
一	2^2	=	4	1^2	=	1
个	1^2	=	1	1^2	=	1
例子	1^2	=	1	1^2	=	1
另	0^2	=	0	0^2	=	0
话题	0^2	=	0	0^2	=	0
			11			5

文本A

这 / 一 / 句 / 话 / 只 / 是 / 一 / 个 / 例子

文本B

这 / 是 / 一 / 个 / 例子

文本C

另 / 一 / 个 / 话题

3. 技术：文本向量的相似度计算

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A}\cdot\mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum\limits_{i=1}^{n} A_i B_i}{\sqrt{\sum\limits_{i=1}^{n} A_i^2}\sqrt{\sum\limits_{i=1}^{n} B_i^2}}$$

常用的向量相似度计算方法：cosine similarity

词汇	向量A	向量B	向量C
这	1	1	0
句	1	0	0
话	1	0	0
只	1	0	0
是	1	1	0
一	2	1	1
个	1	1	1
例子	1	1	0
另	0	0	1
话题	0	0	1

sim(A,B) =$\frac{6}{\sqrt{11}\sqrt{5}}$

= 0.809

sim(A,C) =$\frac{3}{\sqrt{11}\sqrt{4}}$

= 0.452

文本B比起文本C更相似于文本A

文本A

这 / 一 / 句 / 话 / 只 / 是 / 一 / 个 / 例子

文本B

这 / 是 / 一 / 个 / 例子

文本C

另 / 一 / 个 / 话题

3. 技术：Cosine similarity和向量空间

Document 1: cat cat cat cat

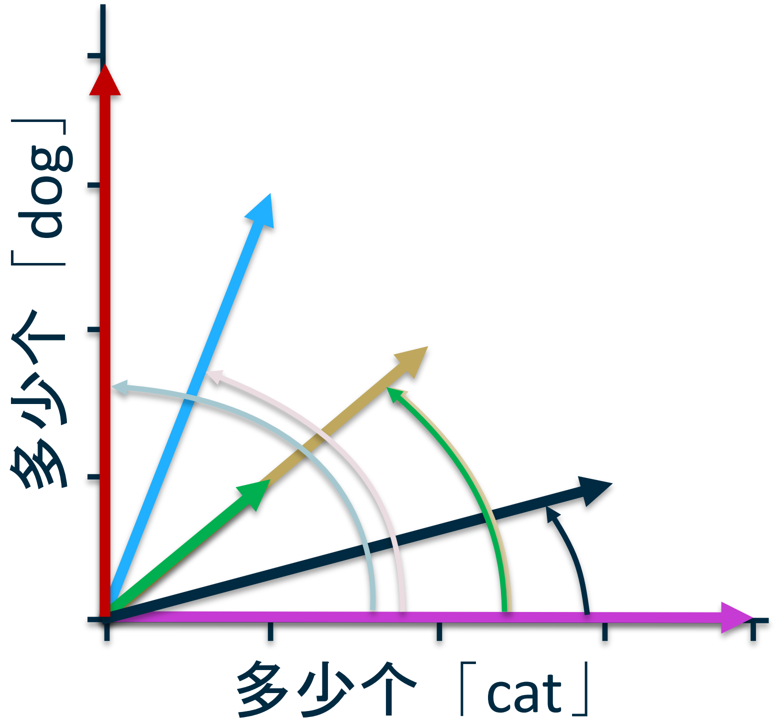
Document 2: cat dog

Document 3: dog dog cat dog

Document 4: dog dog cat cat

Document 5: cat cat cat dog

Document 6: dog dog dog dog



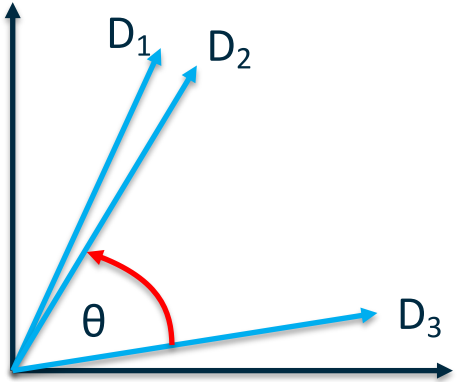
- Document 5 is the closest document to document 1
- Document 2 and document 4 are “the same”

3. 技术：文本向量的相似度计算

另一个常用技术：IDF（Inverse Document Frequency）

- TF向量中，每一个词汇中的单词的重要性是一样的
  - 例如「一」、「的」都看作和「例子」、「北大」、「历史系」一样重要
- 通常反而是越罕见的词汇约重要
  - 例如：两个文本都多次出现「一」和「的」并不代表它的内容相关；两个文本都有许多「北大」和「历史系」，内容就有某种相似性
- TF-IDF会把出现在语料库中越少数文本中的词汇视为越重要

3. Cosine similarity



How similar are two documents, e.g. S_1 and S_2?

- Compare their vectors:

Cosine similarity for vectors A and B:

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

$\Rightarrow$ Cosine similarity is a number from 0 to 1, and:

- 0 when A and B are orthogonal
  - Intuitively (for TF vectors) S_1 and S_2 have no terms in common at all
- 1 when A and B are scalar multiples of one another
  - Intuitively (for TF), all terms in S_1 and S_2 occur in identical proportions

3. Inverse Document Frequency (IDF)

So far, all shared tokens between S_1 and S_2 count *equally* towards similarity

- Includes common terms like “the”, “a”, and punctuation

Intuitively, S_1 and S_2 sharing “persuasion” is more significant than sharing “,”

- [“,” in fact contributes *much more* to sim(S_2,S_3), because it appears twice]

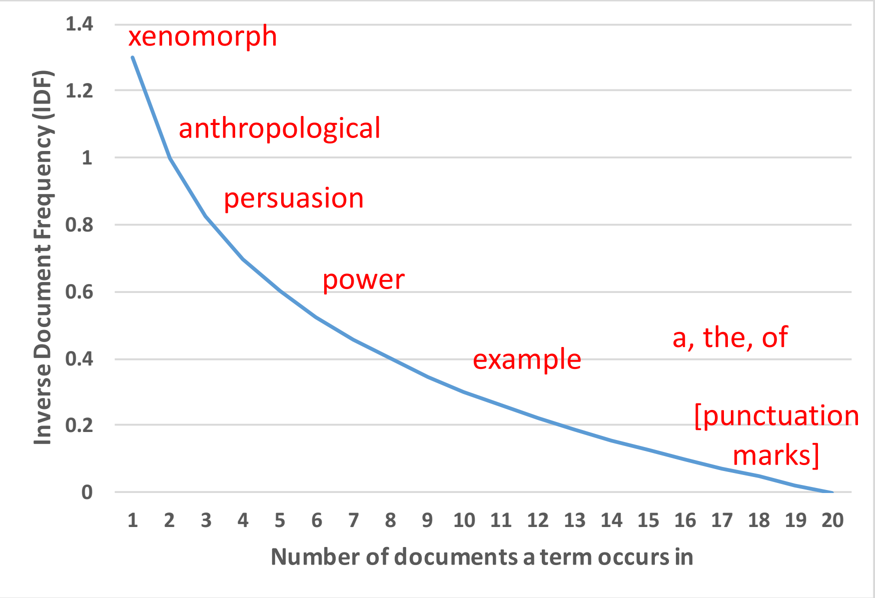
Easy approaches:

- Remove punctuation
- Remove common or “meaningless” words (“the”, “a”, etc.) – “stop words”
  - Both are language-specific: rely on some knowledge of the language
  - 。 ， 、 《 》 「 」 『 』 ？ ！ “ ” vs . , : ! ? “” " [ ] { } …

3. Inverse Document Frequency (IDF)

Heuristic technique (mainly for larger corpora than our example)

- Terms occurring in *many* (or all) documents provide less information
  - Co-occurrence of t in S_1, S_2 more significant for *rare* t than *common* t
  - DF_t = # of documents containing t
  - $IDF_t = \log(N/DF_t)$, N=# of documents
- Replace TF in vectors with $TF*IDF_t$
- "TF-IDF" vectors



Example of IDF in a corpus: illustration only, not calculated from our earlier example

3. Document vectors

Usually very sparse

- Most components of most vectors are 0
  - Most documents do not contain most vocabulary

Useful where “bag of words” assumption is reasonable

- E.g. topic modelling: interested in “content” in the sense of subject matter
  - Order and context less important than frequent mentions of terms
- vs e.g. text reuse => vectors can only be used to detect reused *vocabulary*

3. Cosine similarity – example

Two versions of “the same” novel

Substantial differences

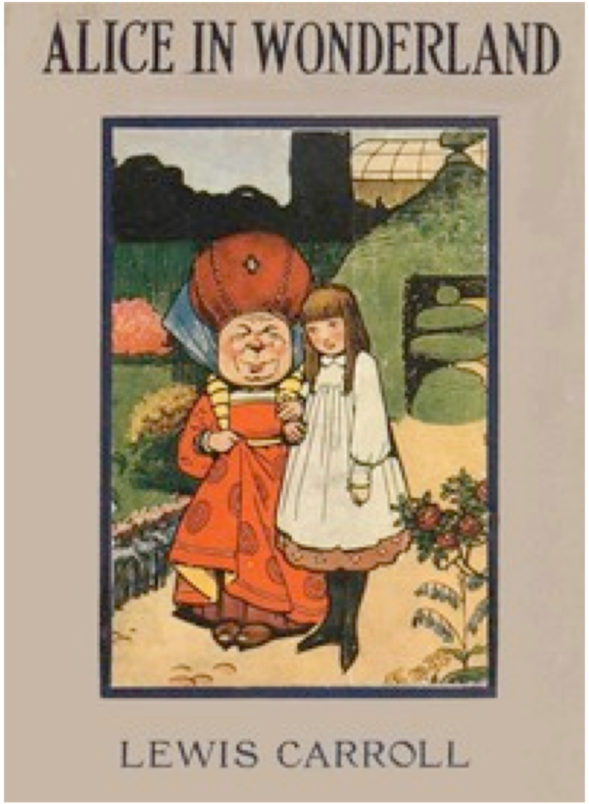
Different chapters

Different lengths

Some parts very similar

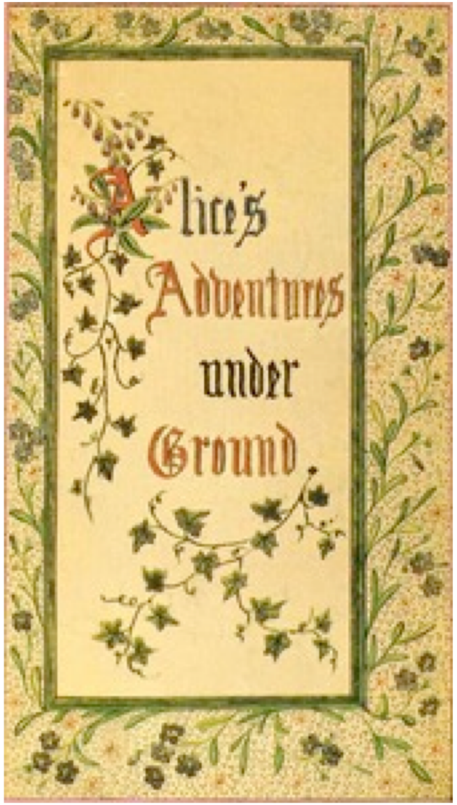
Some parts unique to each

a)



12 chapters, 1865

b)



4 chapters, 1864

3. Cosine similarity – example

Similarity matrix:

Each cell colored by $\text{sim}_{\text{cosine}}(D_i, D_j)$

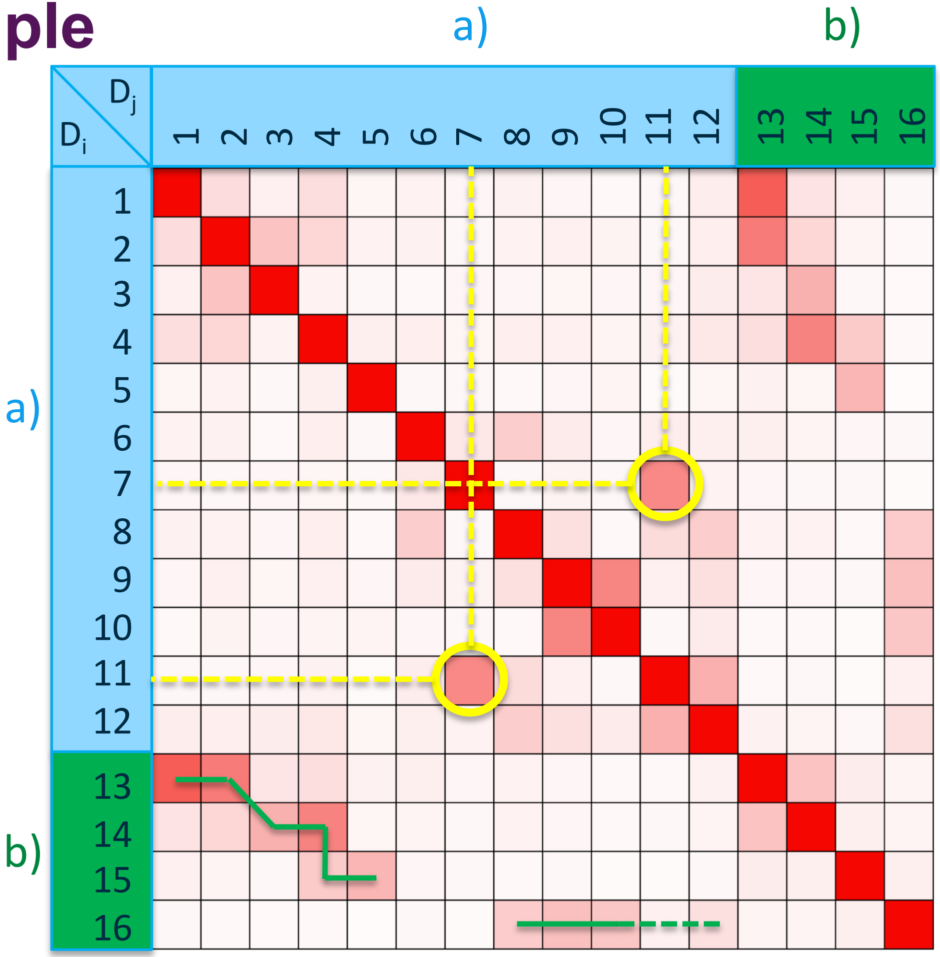
Solid white: 0 solid red: 1

Blue & green two document groups

a) Alice in Wonderland

b) Alice's Adventures Underground

Green line indicates actual reorganization of texts a) and b)



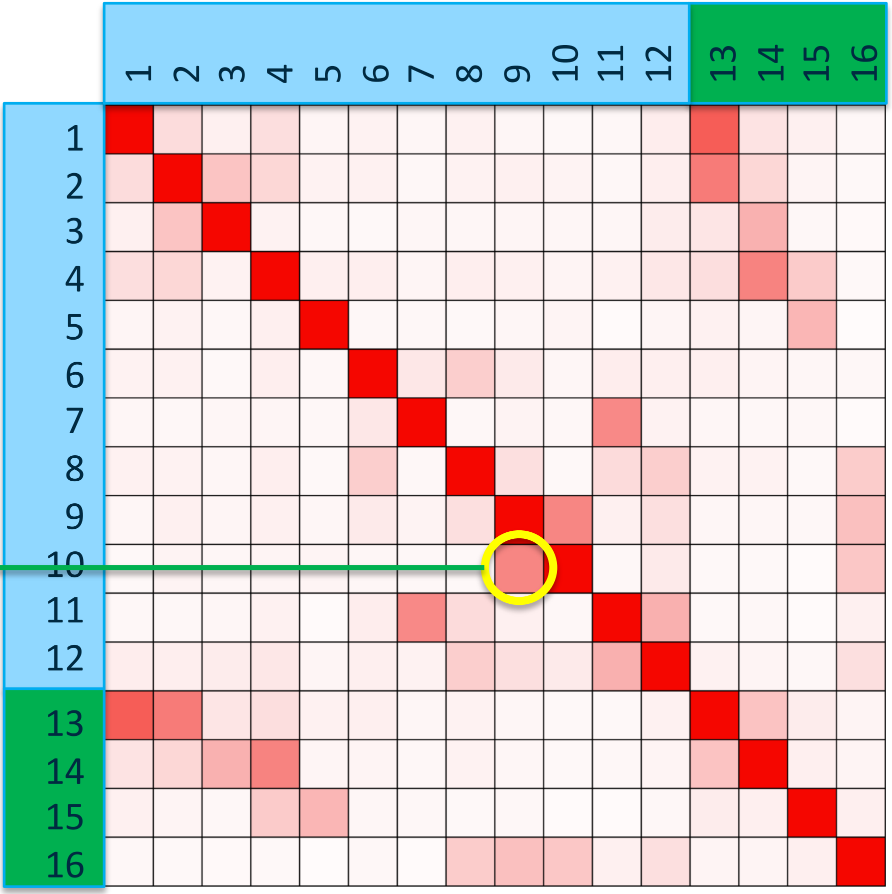
Durham University

3. Cosine similarity – example Interpretability

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A}\cdot\mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\boxed{\sum_{i=1}^{n} A_i B_i}}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

Vocabulary sorted by A_iB_i

t_i	A_i	B_i	A_iB_i
Turtle	0.381	0.47	0.179
Mock	0.368	0.435	0.16
Gryphon	0.342	0.292	0.0999
Soup	0.175	0.042	0.0074
sea	0.0548	0.0311	0.0017
course	0.0328	0.0435	0.0014
school	0.0159	0.084	0.0013
Tis	0.0394	0.0261	0.001
replied	0.0312	0.0236	0.0007
Thank	0.0317	0.021	0.0007
old	0.0131	0.0435	0.0006
different	0.0372	0.0123	0.0005



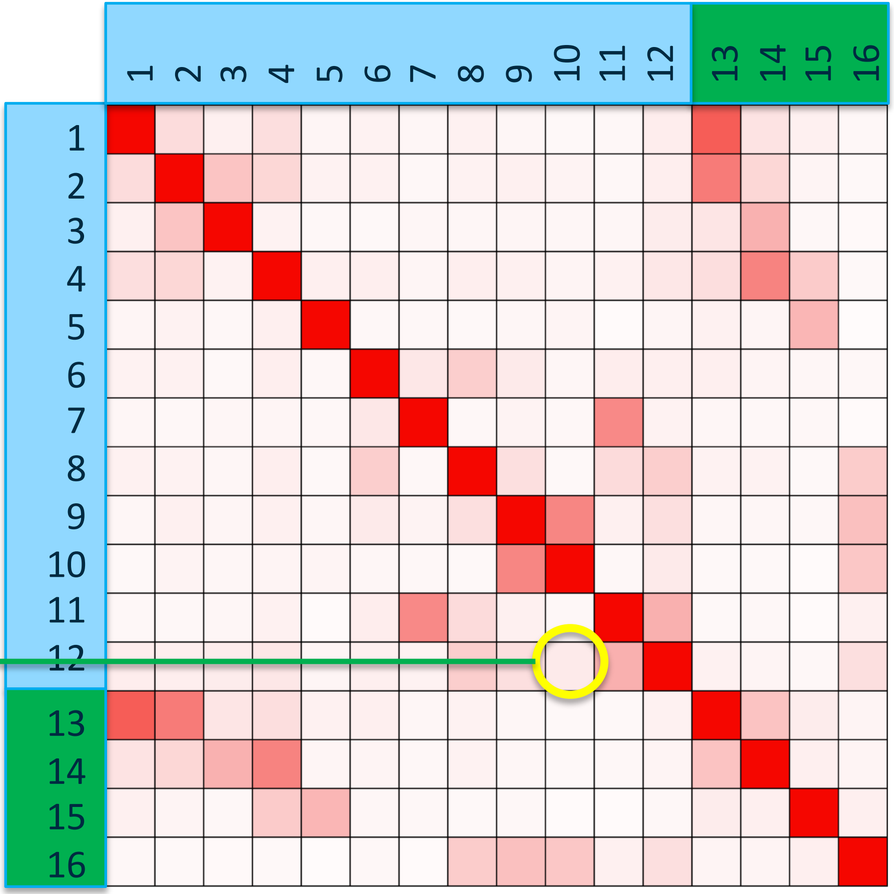
3. Cosine similarity – example Interpretability

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\boxed{\sum_{i=1}^{n} A_i B_i}}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

t_i	A_i	B_i	A_iB_i
—	0.132	0.123	0.0162
Turtle	0.378	0.0425	0.0161
Mock	0.365	0.0425	0.0155
Gryphon	0.339	0.0356	0.0121
sobs	0.0315	0.0513	0.0016
course	0.0326	0.0318	0.001
around	0.0157	0.0513	0.0008
Rabbit	0.0078	0.101	0.0008
White	0.0078	0.101	0.0008
told	0.0219	0.0356	0.0008
trial	0.0219	0.0356	0.0008
him	0.0141	0.0516	0.0007

In this case, the ***most*** *significant contribution is from noise!*

Normalization and data cleaning are important!



3. 技术：n-grams（动机）

词袋模型的明显缺点：完全忽略顺序！

In a bag of words representation (e.g. TF-IDF vectors), these two documents have *identical* representations!

positive

D_1 I usually hate scifi but I love this one.

D_2 I usually love scifi but I hate this one.

negative

$$\begin{array}{ccl} D_1 & D_2 & \\ \left[\begin{array}{c} 2 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{array}\right] & \left[\begin{array}{c} 2 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{array}\right] & \begin{array}{l} \text{I} \\ \text{usually} \\ \text{hate} \\ \text{scifi} \\ \text{but} \\ \text{love} \\ \text{this} \\ \text{one} \end{array} \end{array}$$

關雎 /， / 樂 / 而 / 不 / 淫 /， / 哀 / 而 / 不 / 傷 /。

關雎 /， / 傷 / 而 / 不 /哀 /， / 淫 / 而 / 不 / 樂 /。

3. 技术：n-grams

什么是一个「n-gram」？

- 「n」是一个固定数字（例如，n=1；n=2；n=3；…）
- 一个「n-gram」是指n个token按顺序出现在一起
  - 1-gram = 一个词/字（token）
  - 2-gram = 两个词/字
  - 3-gram = 三个词/字
  - …

3. 技术：n-grams，n=1 ，以字为单位（token）

天命之謂性，率性之謂道，修道之謂教。

1-gram：哪些词（字）出现在我们的文本？总共出现多少次？

天	1
命	1
之	3
謂	3
性	2
，	2
率	1
道	2
修	1
教	1
。	1

3. 技术：n-grams，n=2，以字为单位（token）

天命之謂性，率性之謂道，修道之謂教。

2-gram：哪些长度等于2的词（字）串出现在我们的文本？总共出现多少次？

天命	1
命之	1
之謂	3
謂性	1
性，	1
，率	1
率性	1
性之	1
謂道	1
道，	1
，修	1
修道	1
道之	1
謂教	1
教。	1

3. 技术：n-grams，n=3，以字为单位（token）

天命之謂性，率性之謂道，修道之謂教。

3-gram：哪些长度等于3的词（字）串出现在我们的文本？总共出现多少次？

天命之	1
命之謂	1
之謂性	1
謂性，	1
性，率	1
，率性	1
率性之	1
性之謂	1
之謂道	1
謂道，	1
道，修	1
，修道	1
修道之	1
道之謂	1
之謂教	1
謂教。	1

3. 技术：n-grams

天下	1218	而不	382	君子	186
諸侯	964	天下	372	之子	62
將軍	865	人之	300	我心	46
以為	846	君子	299	子之	42
於是	843	也故	240	文王	41
太子	824	之謂	220	不我	38
二十	615	之所	217	不可	37
天子	608	者也	197	見君	33
《史記》		《荀子》		《詩經》	

3. 技术：正规化

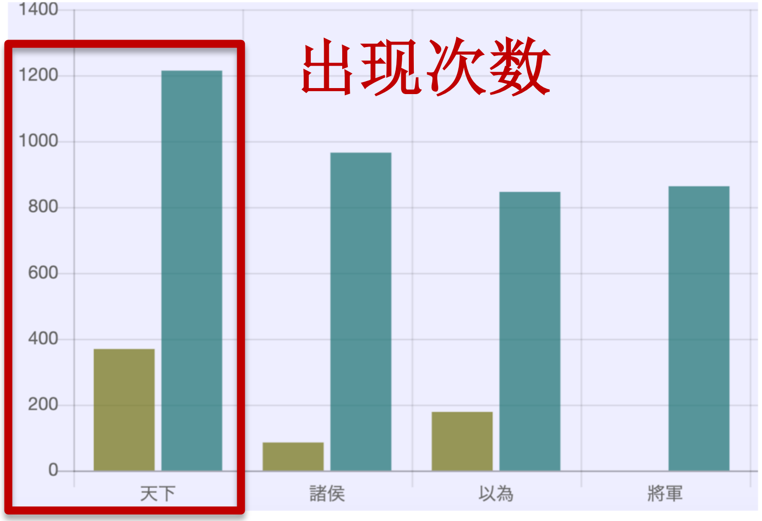
天下	1218
諸侯	964
將軍	865
以為	846
於是	843

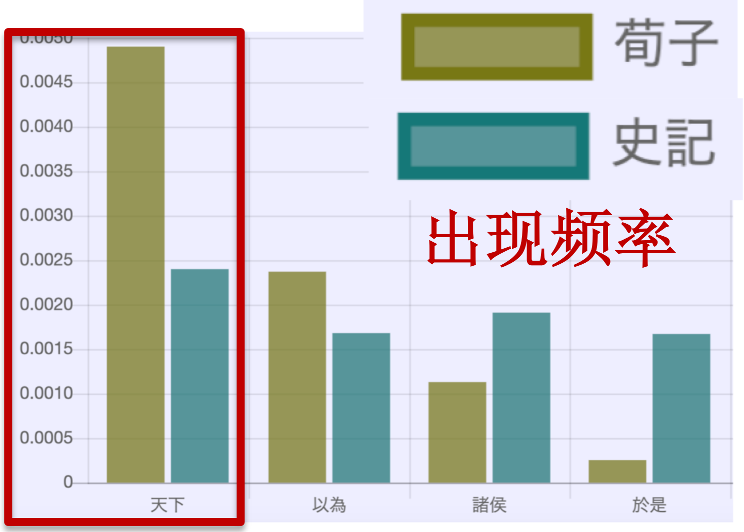
《史記》
>50万字

而不	382
天下	372
人之	300
君子	299
也故	240

《荀子》
7.5万字

哪一部「天下」用的多？





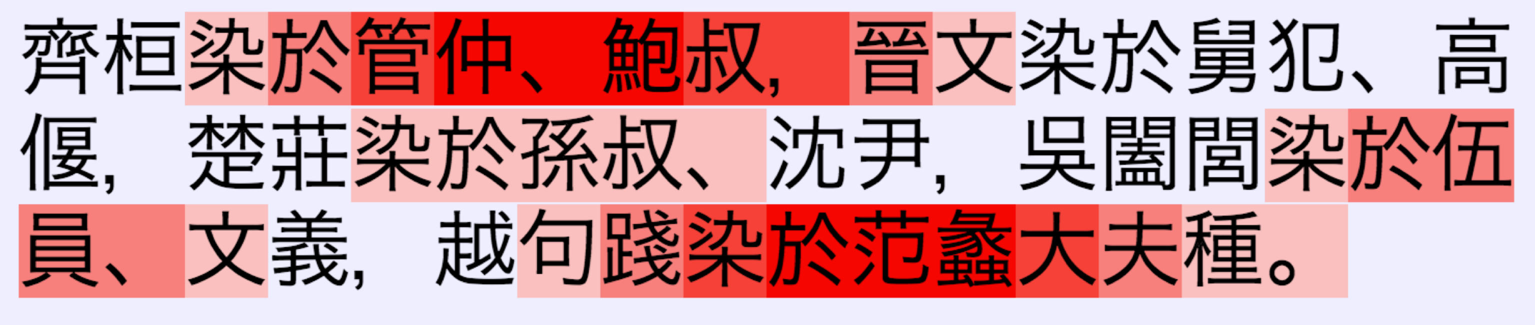
Durham University

3. 技术：用n-gram找出互文性

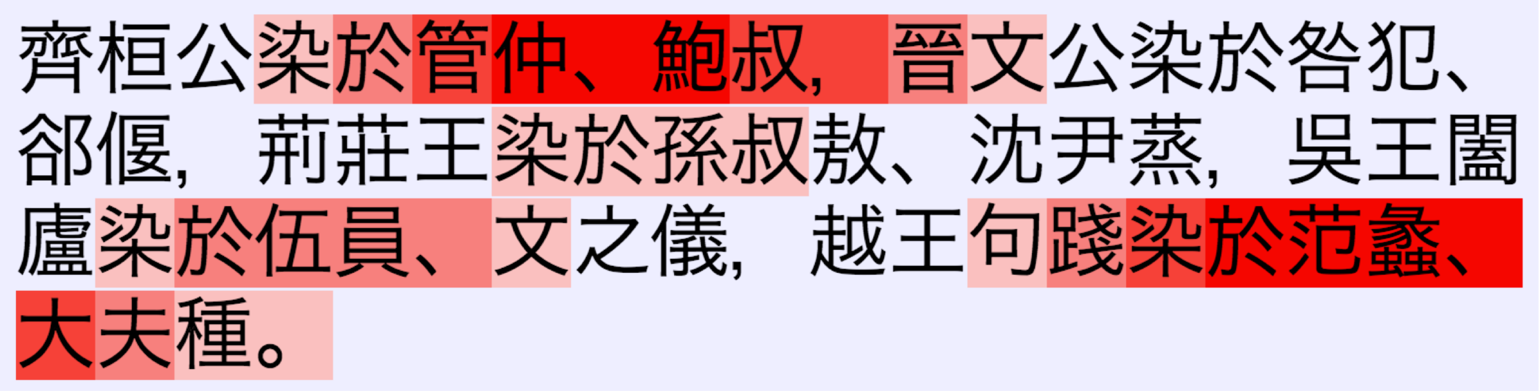
- Example: n=4
- 齊桓染於管仲、鮑叔，晉文染於舅犯、高偃，楚莊染於孫叔、沈尹，吳闔閭染於伍員、文義，越句踐染於范蠡大夫種。
- 齊桓公染於管仲、鮑叔，晉文公染於咎犯、郤偃，荊莊王染於孫叔敖、沈尹蒸，吳王闔廬染於伍員、文之儀，越王句踐染於范蠡、大夫種，

3. 技术：用n-gram找出互文性

- Example: n=4



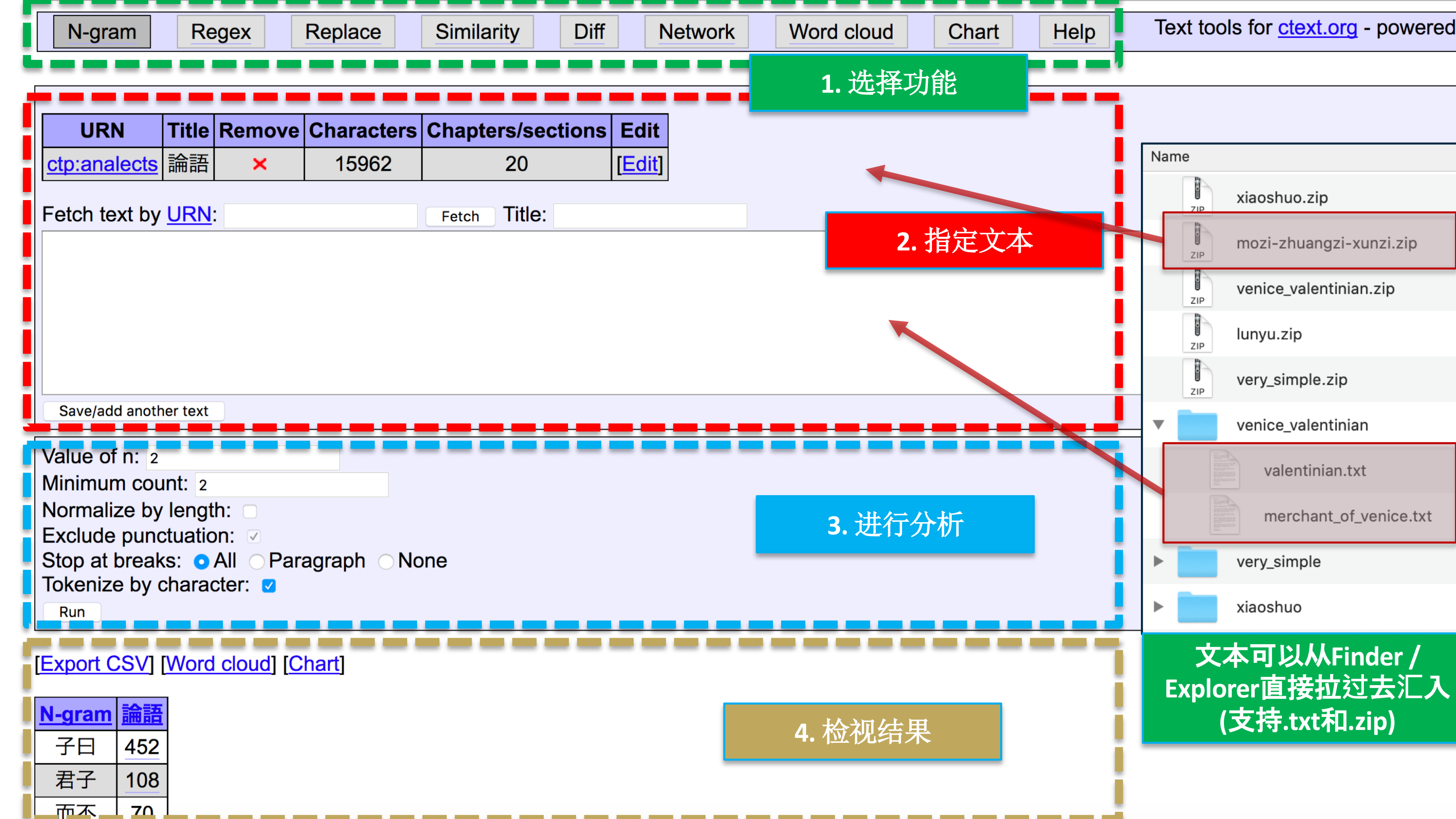
齊桓染於管仲、鮑叔，晉文染於舅犯、高偃，楚莊染於孫叔、沈尹，吳闔閭染於伍員、文義，越句踐染於范蠡大夫種。



齊桓公染於管仲、鮑叔，晉文公染於咎犯、郤偃，荊莊王染於孫叔敖、沈尹蒸，吳王闔廬染於伍員、文之儀，越王句踐染於范蠡、大夫種。

4. 实践

https://text.tools/ctext



1. 选择功能

URN	Title	Remove	Characters	Chapters/sections	Edit
ctp:analects	論語	×	15962	20	[Edit]

Fetch text by URN: Fetch Title:

Save/add another text

2. 指定文本

Name
- xiaoshuo.zip
- mozi-zhuangzi-xunzi.zip
- venice_valentinian.zip
- lunyu.zip
- very_simple.zip
- venice_valentinian
  - valentinian.txt
  - merchant_of_venice.txt
- very_simple
- xiaoshuo

Value of n: 2

Minimum count: 2

Normalize by length: □

Exclude punctuation: ☑

Stop at breaks: ● All ○ Paragraph ○ None

Tokenize by character: ☑

Run

3. 进行分析

[Export CSV] [Word cloud] [Chart]

N-gram	論語
子曰	452
君子	108
而不	70

4. 检视结果

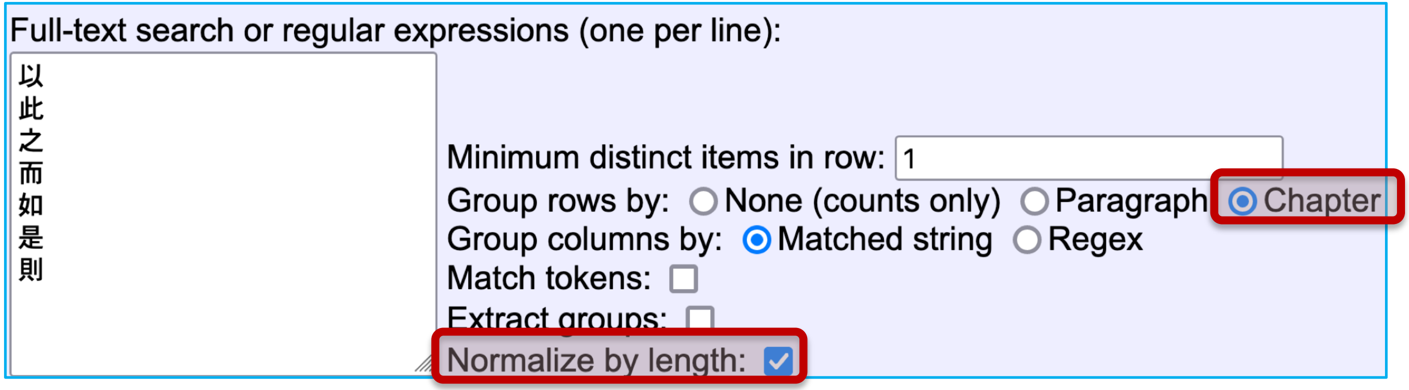
文本可以从Finder / Explorer直接拉过去汇入 (支持.txt和.zip)

4. 实践：推荐路线

基本技术	试用文本	主要操作方式
Regex	論語	Regex => 在「Full-text search」輸入regex内容，按「Run」
N-gram	墨子+莊子+荀子	Similarity => n=7 => Run => 点击红色部分； 点「Similarity matrix」=> 「Toggle values」； 点「Chapter summary」=> 「Create graph」 => 「Draw」； 双击任网络图中的边
Cosine	墨子	Vectors => Run => Toggle values => 点其中有红色的方块； 点其中的字词现实语料库中的分布； 点图中的内容以现实对应内容； 点 「Vectors」回到原本的位置

4. 实践：推荐路线

基本技术	试用文本	主要操作方式
PCA	明代小说	Regex => 在「Full-text search」輸入regex：一行一个regex（如下），选择「Chapter」和「Normalize by length」，点「Run」 再点「Summary」 => 「Create vectors」 => 「Run PCA」



Full-text search or regular expressions (one per line):

以
此
之
而
如
是
則

Minimum distinct items in row: 1

Group rows by: ○ None (counts only) ○ Paragraph ● Chapter

Group columns by: ● Matched string ○ Regex

Match tokens: □

Extract groups: □

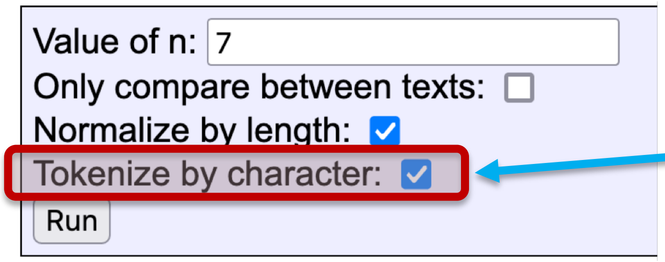
Normalize by length: ☑

4. 实践：推荐路线

基本技术	试用文本	主要操作方式
PCA + TF-IDF（所有词汇）	明代小说	Vectors => Run => Run PCA 有什么样的结果？结果跟文体分析/作者辨别有关吗？

4. 实践：推荐路线

基本技术	试用文本	主要操作方式
分词	英文文本/ 现代汉语 / 已分词文本	Transform => 在「Register」 => 选择转换类型 => Run / Apply to all 以进行分词



Value of n: 7

Only compare between texts: ☐

Normalize by length: ☑

Tokenize by character: ☑

注意：用所有英文例子/以分词的文本时，要吧「Tokenize by character」取消；否则，系统会一个字母一个字母处理

文体分析 Stylometry

- 经常用在作者辨别（authorship attribution）
- 文本A的作者有不确定性：最可能是作者X或者作者Y；而且，我们有其它文本确定是作者X和作者Y写的。
- 不同作者有潜意识的写作特征（如：某一些单词、语法结构等用的多）

文体分析：Principal Component Analysis (PCA)

- Modeling texts, we often use *vectors* to represent documents
  - We used these to represent text digitally
  - We used these to identify textual similarity
  - LDA topic modeling is based on the same idea
  - Vectors are convenient for computers…
  - ...But: too many dimensions for us to *visualize*
    - Hard to see patterns, even when clear patterns exist

Documents as Vectors (again!)

Document 1

我们 / 不再 / 简单 / 追求 / 经济 / 增长 / 的 / 高 / 速度 / ， / 而 是 / 强调 / 经济 / 发展 / 的 / 质量 / 和 / 效益 / 。

Document 2

提高 / 实体 / 经济 / 的 / 整体 / 素质 / 和 / 竞争力 / 。

All words	Doc. 1	Doc. 2
我们	1	0
的	2	1
提高	0	1
实体	0	1
和	1	1
不再	1	0
简单	1	0
追求	1	0
经济	2	1
增长	1	0
…	…	…

Durham University

Documents as Vectors (again!)

All words	Doc. 1	Doc. 2
我们	1	0
的	2	1
提高	0	1
实体	0	1
和	1	1
不再	1	0
简单	1	0
追求	1	0
经济	2	1
增长	1	0
	…	…

doc1=[1,2,0,0,1,1,…]

doc2=[0,1,1,1,1,0,…]

If only first 3 words:

doc1=[1,2,0]

doc2=[0,1,1]

If only first 2 words:

doc1=[1,2]

doc2=[0,1]

维度、向量、和可视化

1 Dimension: x-coordinate

Example: [1]

[3]

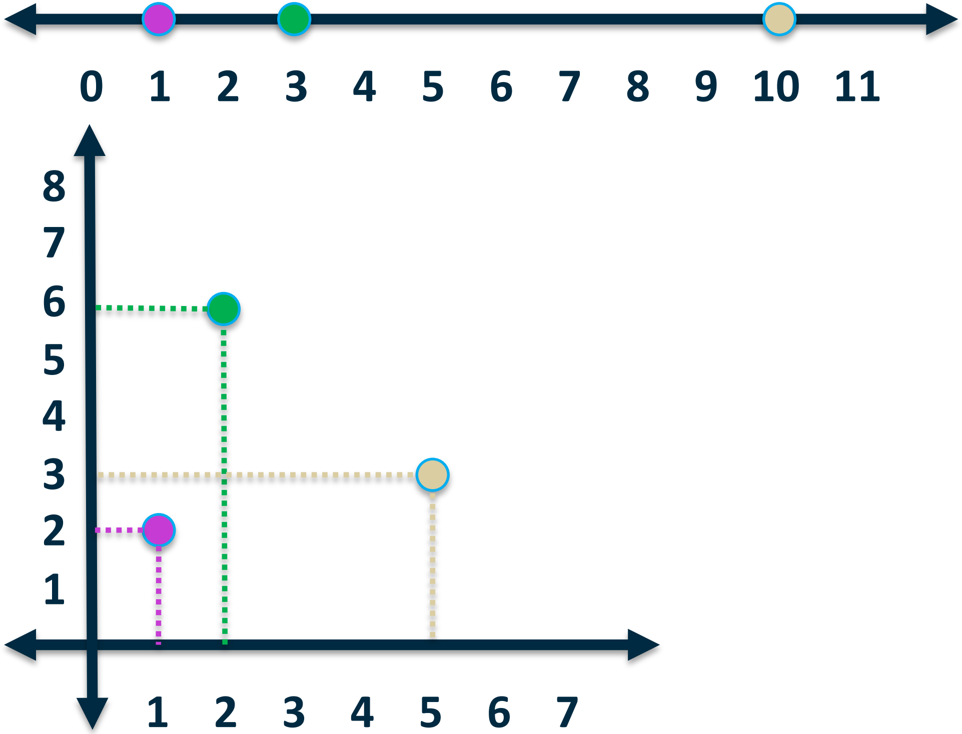
[10]

2 Dimensions: x and y

Example: [1, 2]

[2, 6]

[5, 3]

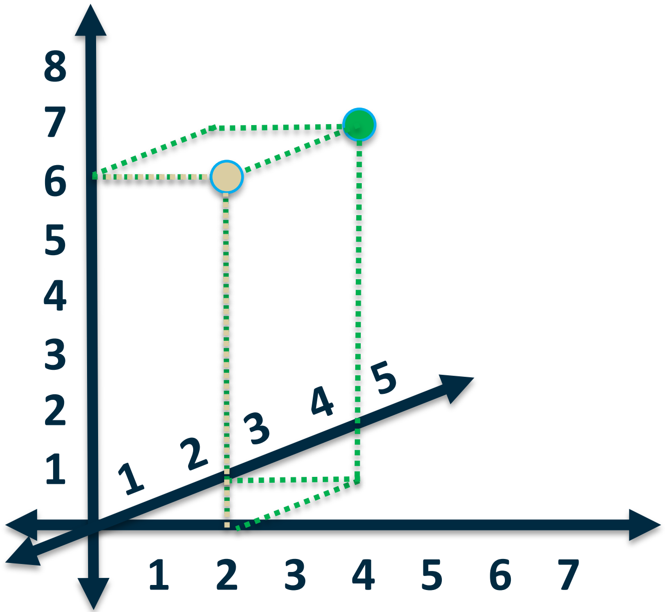


维度、向量、和可视化

3 Dimensions: x, y, and z

Example: [2, 6, 0]

[2, 6, 2]



电脑屏幕是个二维空间！上述的图是一个二维投影（2D projection）。

维度、向量、和可视化

4 Dimensions: x, y, z, a

Example: [1, 2, 5, 1]

[2, 6, 1, 9]

[5, 6, 3, 8]

? ? ?

PCA也可提供一种投影的方法，从任何维度空间到二维空间

它的特点：

- 根据数据点而选择投影的具体方式
- 是一个linear projection
- 会尽可能保留数据点之间的variance（方差）

PCA能做什么？

让我们可以把握高维度空间中的数据点长什么样子

如果我们看到PCA的不同类数据是（linearly）separable，高维度的点中也是

PCA和文体分析：

- 可以指定我们想分析的词汇，作文本向量，然后用PCA可视化
    - 这样可以看出不同文本的整体用词特征
    - 问题：「文体分析」应该用什么样的词汇去分析？
    - 在文体分析和在主体模型正好相反，我们想忽略内容而重视形式
        - 因此，经常会用虚词当作相关词汇

Principal Component Analysis (PCA)

Computes a *linear* transformation from $\mathbb{R}^N \rightarrow \mathbb{R}^N$ calculated from set of data

- Linear transformation such that in the transformed space:
  - Dimension 1 accounts for greatest proportion of variance of datapoints
  - Dimension 2 accounts for next greatest proportion
    - (while being orthogonal to previous dimensions)
  - …
- Keeping the first M<N components gives a projection into $\mathbb{R}^M$
  - Here we use this for visualization of $\mathbb{R}^N$ in $\mathbb{R}^2$
    - Effectively choosing a projection s.t. maximum variance preserved
    - Gives a good sense of e.g. how linearly separable data is

Example – Wizard of Oz

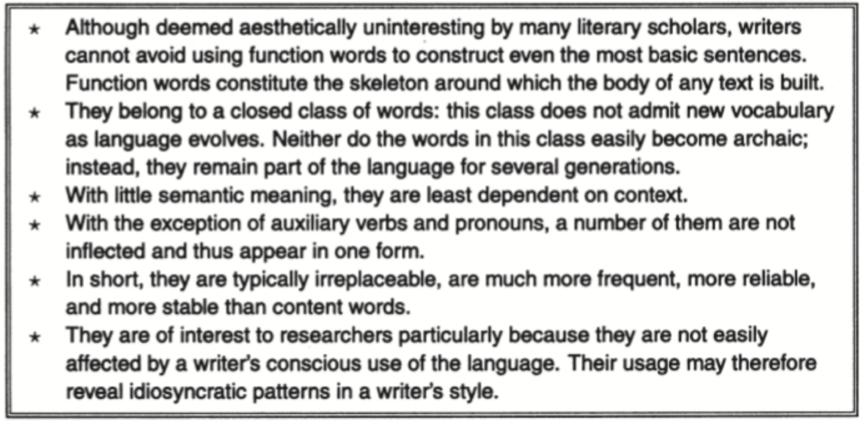
两个作者写了一系列小说中的所有的（或绝大部分）文本

- 作者一：Baum；作者二：Thompson
- “Wizard of Oz” + 几十本续集
  - 全部都有相同题材（小说、幻想fantasy类型，等）
  - 不同文本中有不同人物，但是内容整体相似
- 可是：这系列中的一本不确定是哪位作者写的

[Similar in principle to another well-known example, “The Federalist Papers”]

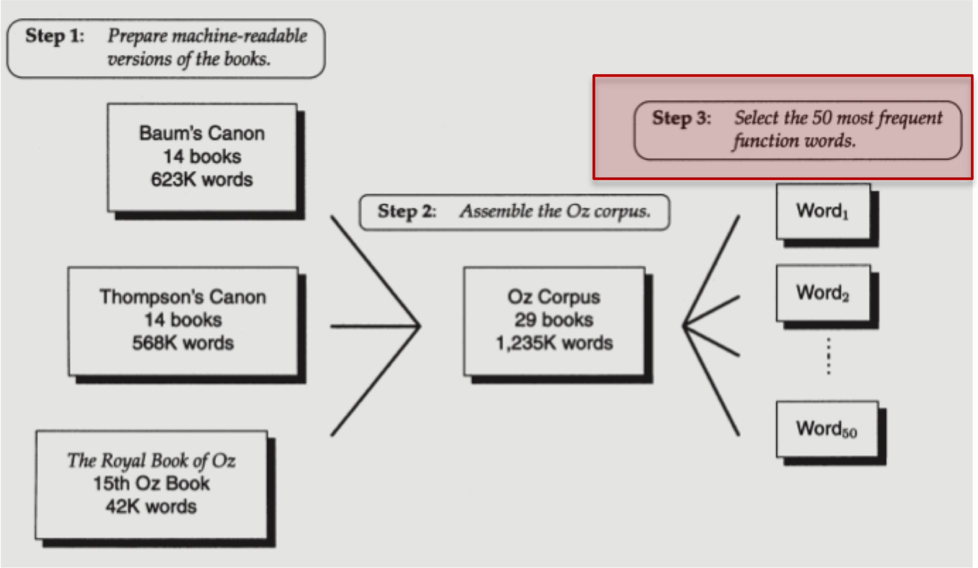
Who Wrote the 15th Book of Oz? An Application of Multivariate Analysis to Authorship Attribution

Example – Wizard of Oz



- Although deemed aesthetically uninteresting by many literary scholars, writers cannot avoid using function words to construct even the most basic sentences. Function words constitute the skeleton around which the body of any text is built.
- They belong to a closed class of words: this class does not admit new vocabulary as language evolves. Neither do the words in this class easily become archaic; instead, they remain part of the language for several generations.
- With little semantic meaning, they are least dependent on context.
- With the exception of auxiliary verbs and pronouns, a number of them are not inflected and thus appear in one form.
- In short, they are typically irreplaceable, are much more frequent, more reliable, and more stable than content words.
- They are of interest to researchers particularly because they are not easily affected by a writer's conscious use of the language. Their usage may therefore reveal idiosyncratic patterns in a writer's style.

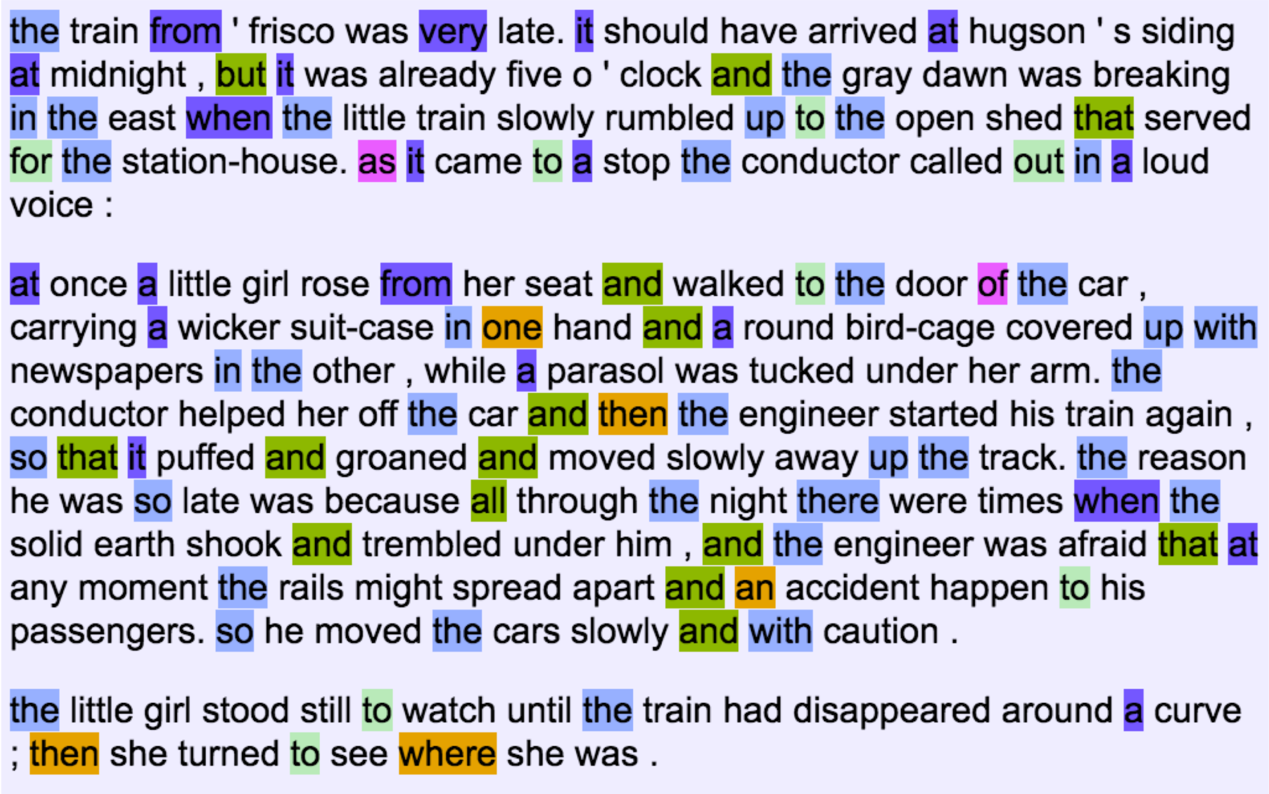
Figure 1. Why function words?



Step 1: *Prepare machine-readable versions of the books.*
Baum's Canon 14 books 623K words
Thompson's Canon 14 books 568K words
The Royal Book of Oz 15th Oz Book 42K words
Step 2: *Assemble the Oz corpus.*
Oz Corpus 29 books 1,235K words
Step 3: *Select the 50 most frequent function words.*
Word$_1$
Word$_2$
Word$_{50}$

[Who Wrote the 15th Book of Oz? An Application of Multivariate Analysis to Authorship Attribution]

Example – Wizard of Oz



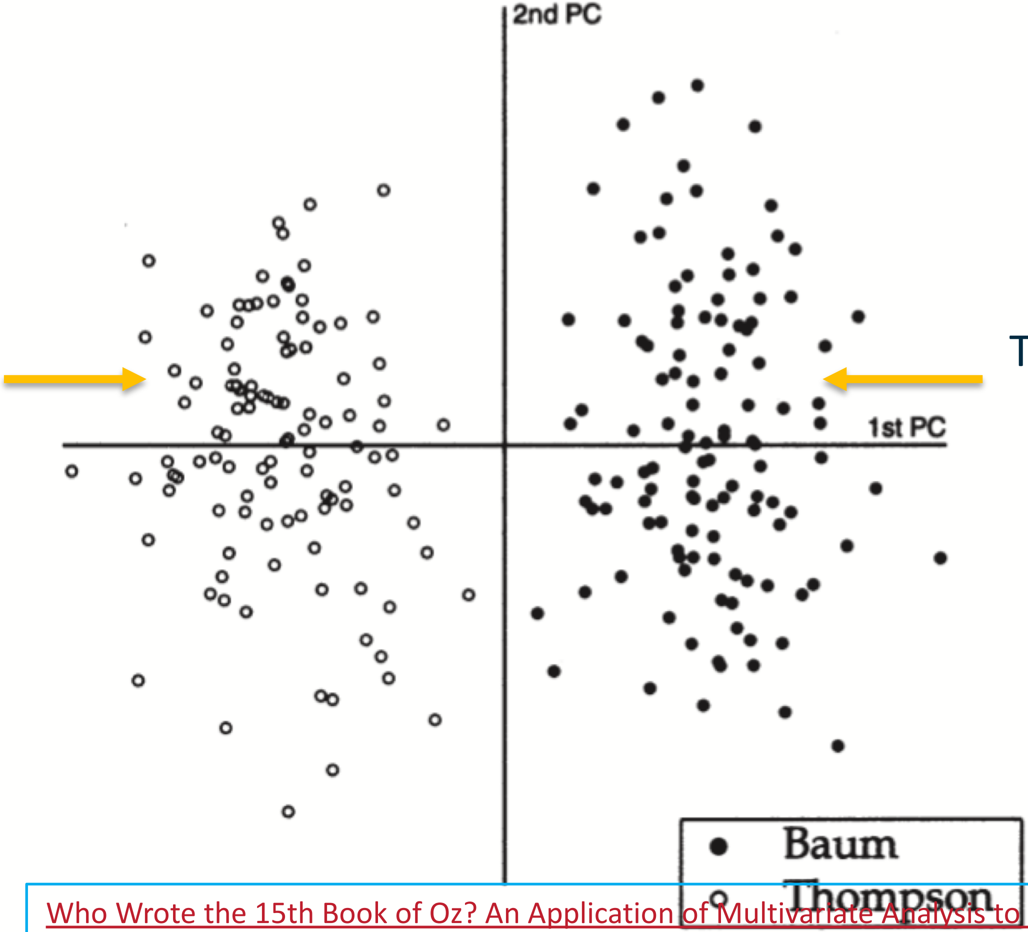
the train from ' frisco was very late. it should have arrived at hugson ' s siding at midnight , but it was already five o ' clock and the gray dawn was breaking in the east when the little train slowly rumbled up to the open shed that served for the station-house. as it came to a stop the conductor called out in a loud voice :

at once a little girl rose from her seat and walked to the door of the car , carrying a wicker suit-case in one hand and a round bird-cage covered up with newspapers in the other , while a parasol was tucked under her arm. the conductor helped her off the car and then the engineer started his train again , so that it puffed and groaned and moved slowly away up the track. the reason he was so late was because all through the night there were times when the solid earth shook and trembled under him , and the engineer was afraid that at any moment the rails might spread apart and an accident happen to his passengers. so he moved the cars slowly and with caution .

the little girl stood still to watch until the train had disappeared around a curve ; then she turned to see where she was .

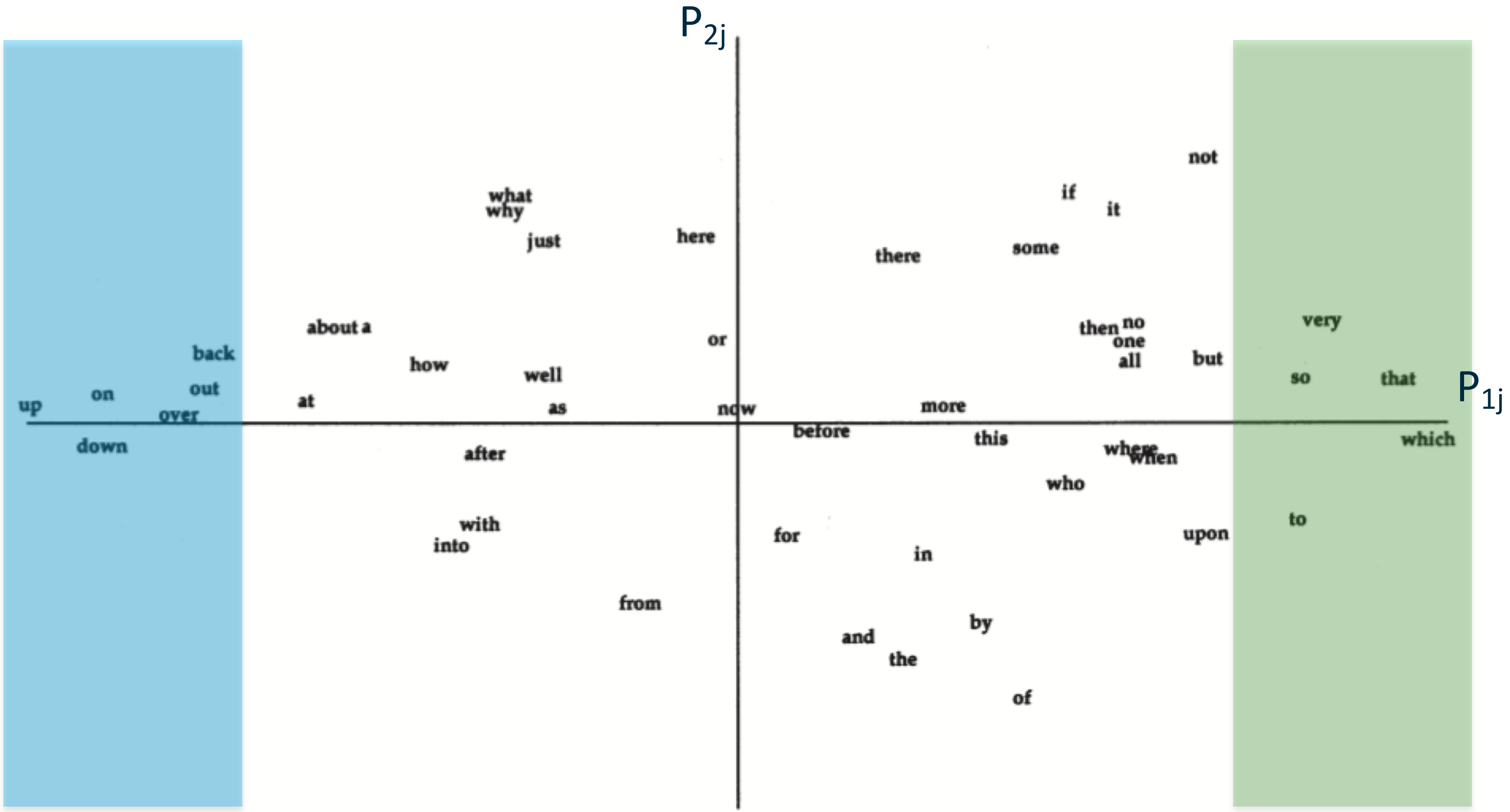
Example -

Texts by Thompson



Texts by Baum

Who Wrote the 15th Book of Oz? An Application of Multivariate Analysis to Authorship Attribution



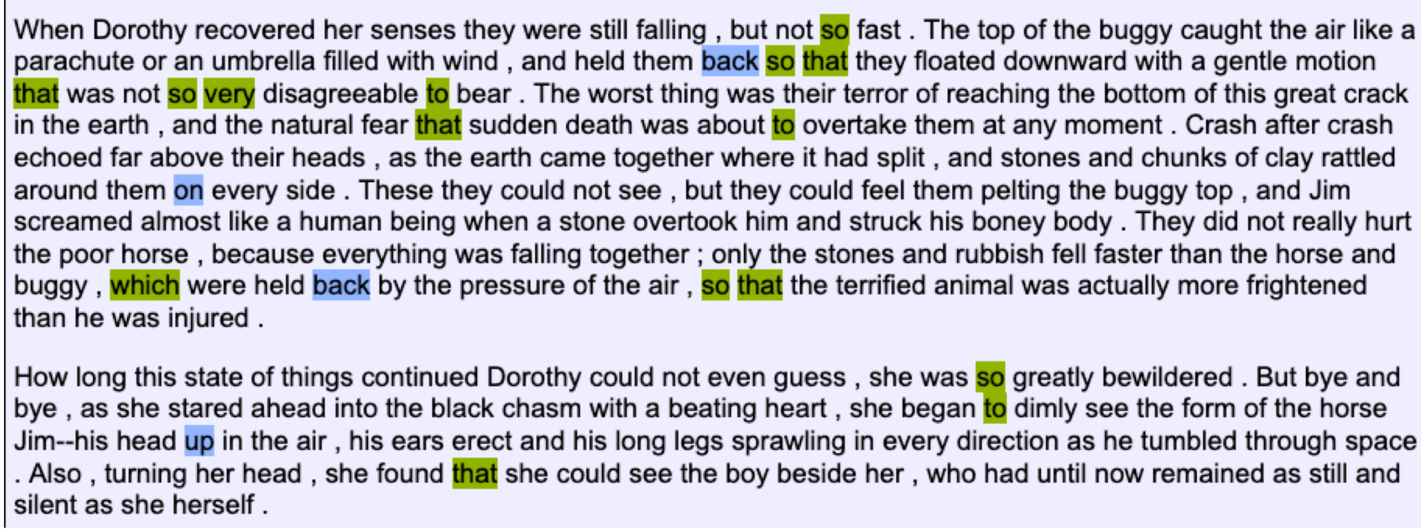
Overrepresented in Thompson

Figure 6. Component loadings.

Overrepresented in Baum

Who Wrote the 15th Book of Oz? An Application of Multivariate Analysis to Authorship Attribution

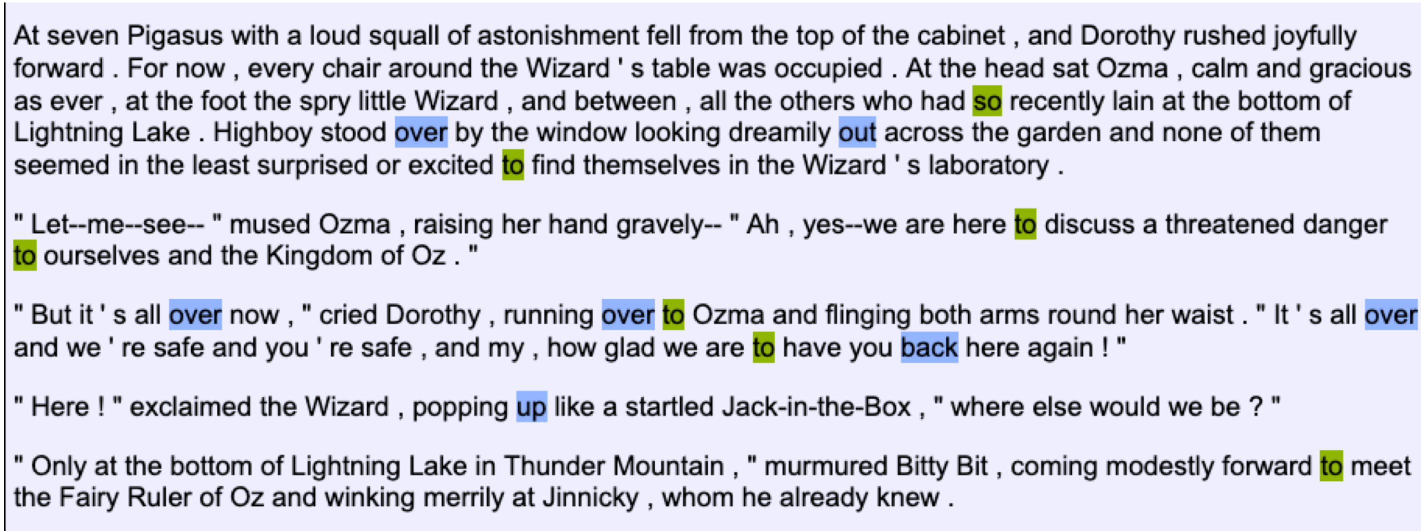
Work by Baum



When Dorothy recovered her senses they were still falling , but not so fast . The top of the buggy caught the air like a parachute or an umbrella filled with wind , and held them back so that they floated downward with a gentle motion that was not so very disagreeable to bear . The worst thing was their terror of reaching the bottom of this great crack in the earth , and the natural fear that sudden death was about to overtake them at any moment . Crash after crash echoed far above their heads , as the earth came together where it had split , and stones and chunks of clay rattled around them on every side . These they could not see , but they could feel them pelting the buggy top , and Jim screamed almost like a human being when a stone overtook him and struck his boney body . They did not really hurt the poor horse , because everything was falling together ; only the stones and rubbish fell faster than the horse and buggy , which were held back by the pressure of the air , so that the terrified animal was actually more frightened than he was injured .

How long this state of things continued Dorothy could not even guess , she was so greatly bewildered . But bye and bye , as she stared ahead into the black chasm with a beating heart , she began to dimly see the form of the horse Jim--his head up in the air , his ears erect and his long legs sprawling in every direction as he tumbled through space . Also , turning her head , she found that she could see the boy beside her , who had until now remained as still and silent as she herself .

Work by Thompson



At seven Pigasus with a loud squall of astonishment fell from the top of the cabinet , and Dorothy rushed joyfully forward . For now , every chair around the Wizard ' s table was occupied . At the head sat Ozma , calm and gracious as ever , at the foot the spry little Wizard , and between , all the others who had so recently lain at the bottom of Lightning Lake . Highboy stood over by the window looking dreamily out across the garden and none of them seemed in the least surprised or excited to find themselves in the Wizard ' s laboratory .

" Let--me--see-- " mused Ozma , raising her hand gravely-- " Ah , yes--we are here to discuss a threatened danger to ourselves and the Kingdom of Oz . "

" But it ' s all over now , " cried Dorothy , running over to Ozma and flinging both arms round her waist . " It ' s all over and we ' re safe and you ' re safe , and my , how glad we are to have you back here again ! "

" Here ! " exclaimed the Wizard , popping up like a startled Jack-in-the-Box , " where else would we be ? "

" Only at the bottom of Lightning Lake in Thunder Mountain , " murmured Bitty Bit , coming modestly forward to meet the Fairy Ruler of Oz and winking merrily at Jinnicky , whom he already knew .

Additional texts [not in the official Oz series] by Baum

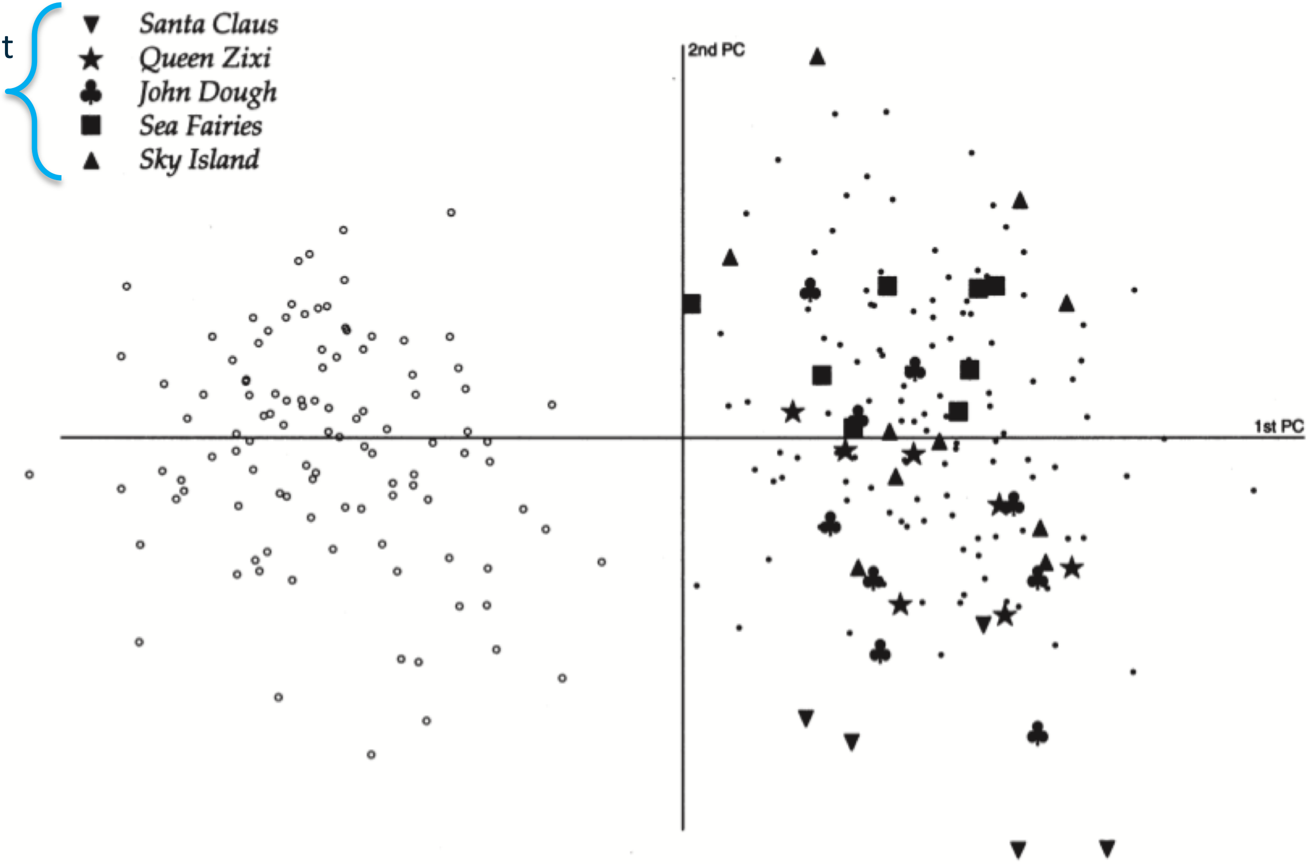


Figure 7. Baum's non-canonical works.

Durham University

[Who Wrote the 15th Book of Oz? An Application of Multivariate Analysis to Authorship Attribution]

Additional texts [other genres] by A) Baum B) Thompson

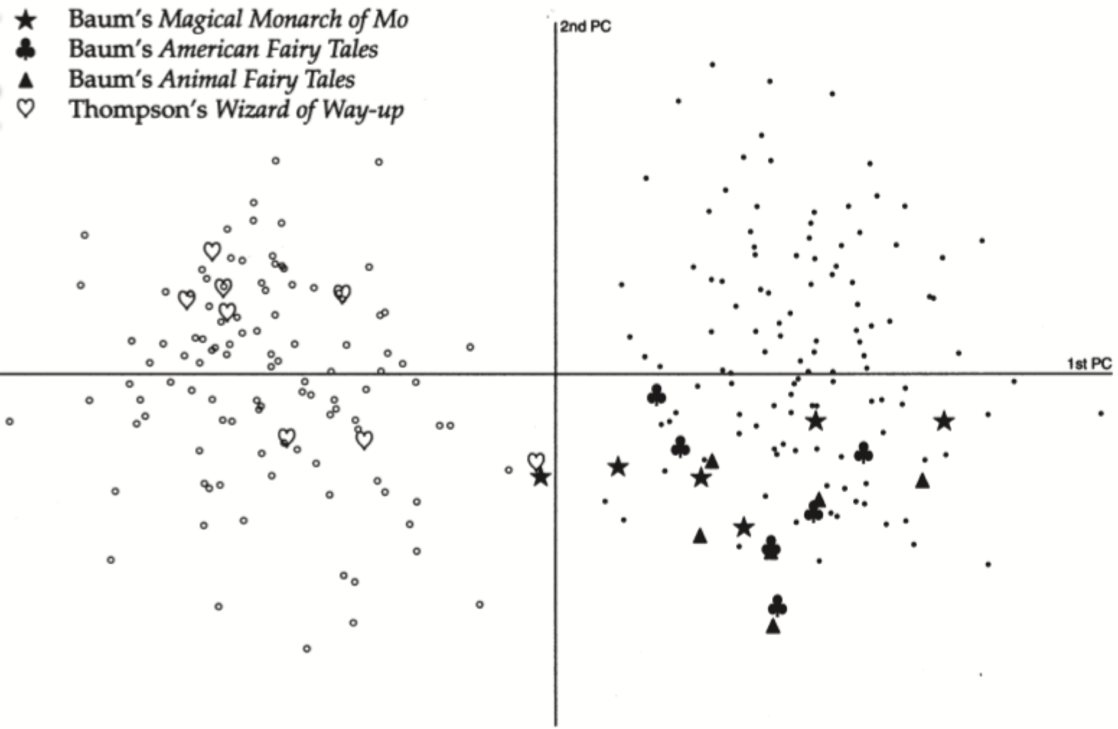


Figure 9. Baum's and Thompson's short stories.

Who Wrote the 15th Book of Oz? An Application of Multivariate Analysis to Authorship Attribution

The contested book

The last book by Baum, *Glinda of Oz*, was published a year after his death. Gardner (Gardner and Nye, p. 40) reports that during the last year and a half of his life, Baum wrote a rough draft of this work. One of his sons edited the published version.

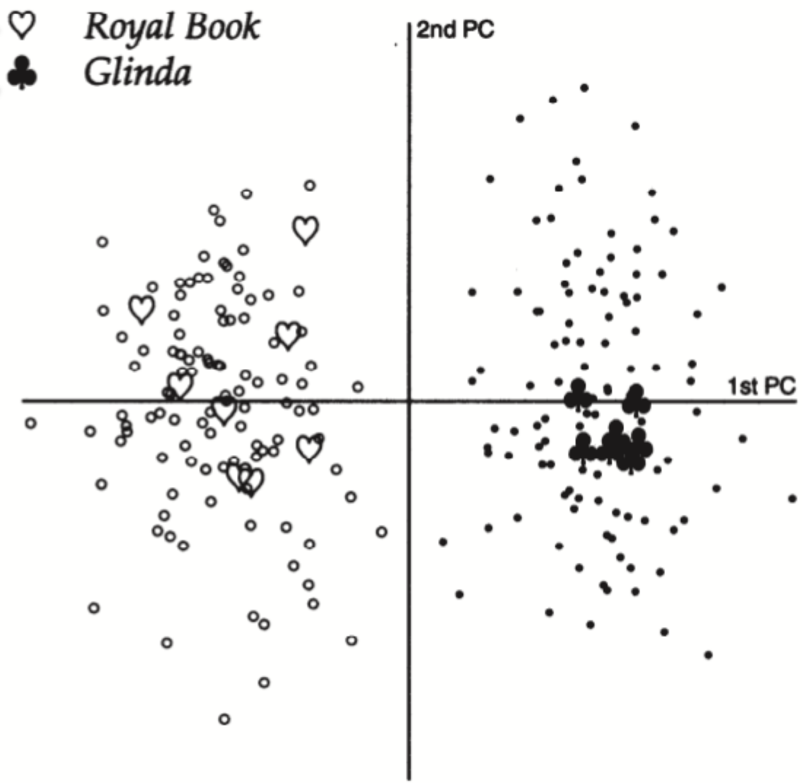


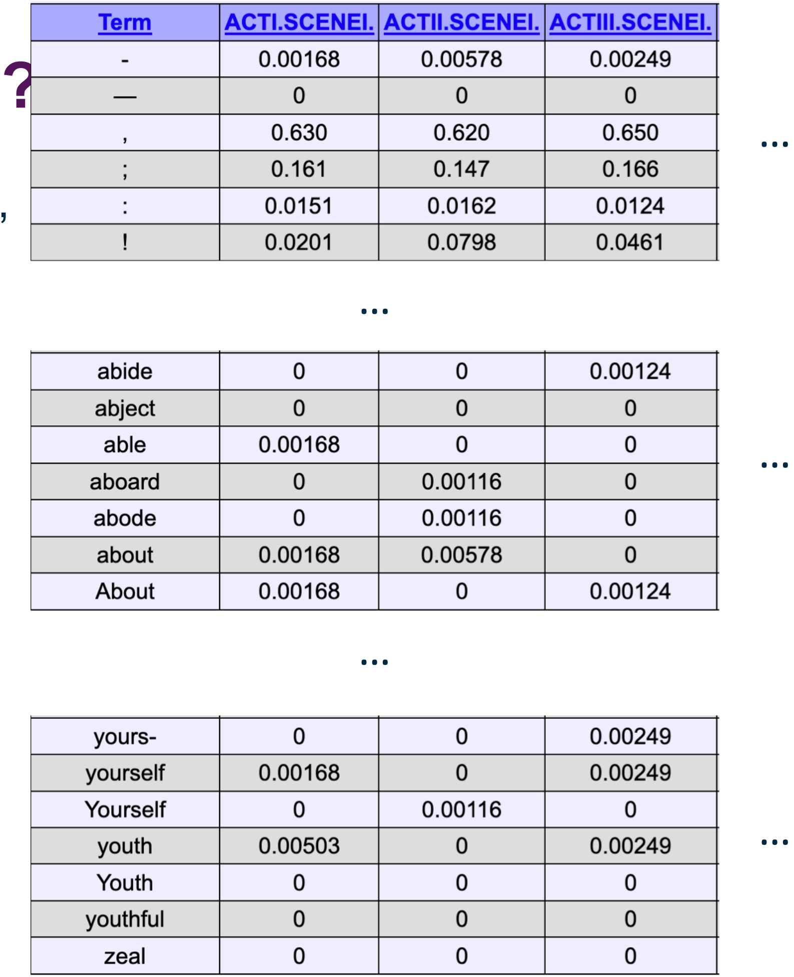
Figure 10. ***The Royal Book of Oz*** **(1921).**

Who Wrote the 15th Book of Oz? An Application of Multivariate Analysis to Authorship Attribution

PCA & Authorship
Can we use “everything”?

- Compare “Valentinian” (by Fletcher) with “Merchant of Venice” (by Shakespeare)
- 1 document per chapter of text
- TF vectors (normalized by dividing by document length)
- |V| dimensions
  - Here |V| ≈ 6000

N.B. As you can see from the tokens shown by the vectors on this slide, this is a simple “back-of-the-envelope” example in which I have *not* been careful about many details (e.g. text preprocessing) that we have already identified as potentially problematic. In your reports, you need to be more careful than this!



Term	ACTI.SCENEI.	ACTII.SCENEI.	ACTIII.SCENEI.
-	0.00168	0.00578	0.00249
—	0	0	0
,	0.630	0.620	0.650
;	0.161	0.147	0.166
:	0.0151	0.0162	0.0124
!	0.0201	0.0798	0.0461

...

...

abide	0	0	0.00124
abject	0	0	0
able	0.00168	0	0
aboard	0	0.00116	0
abode	0	0.00116	0
about	0.00168	0.00578	0
About	0.00168	0	0.00124

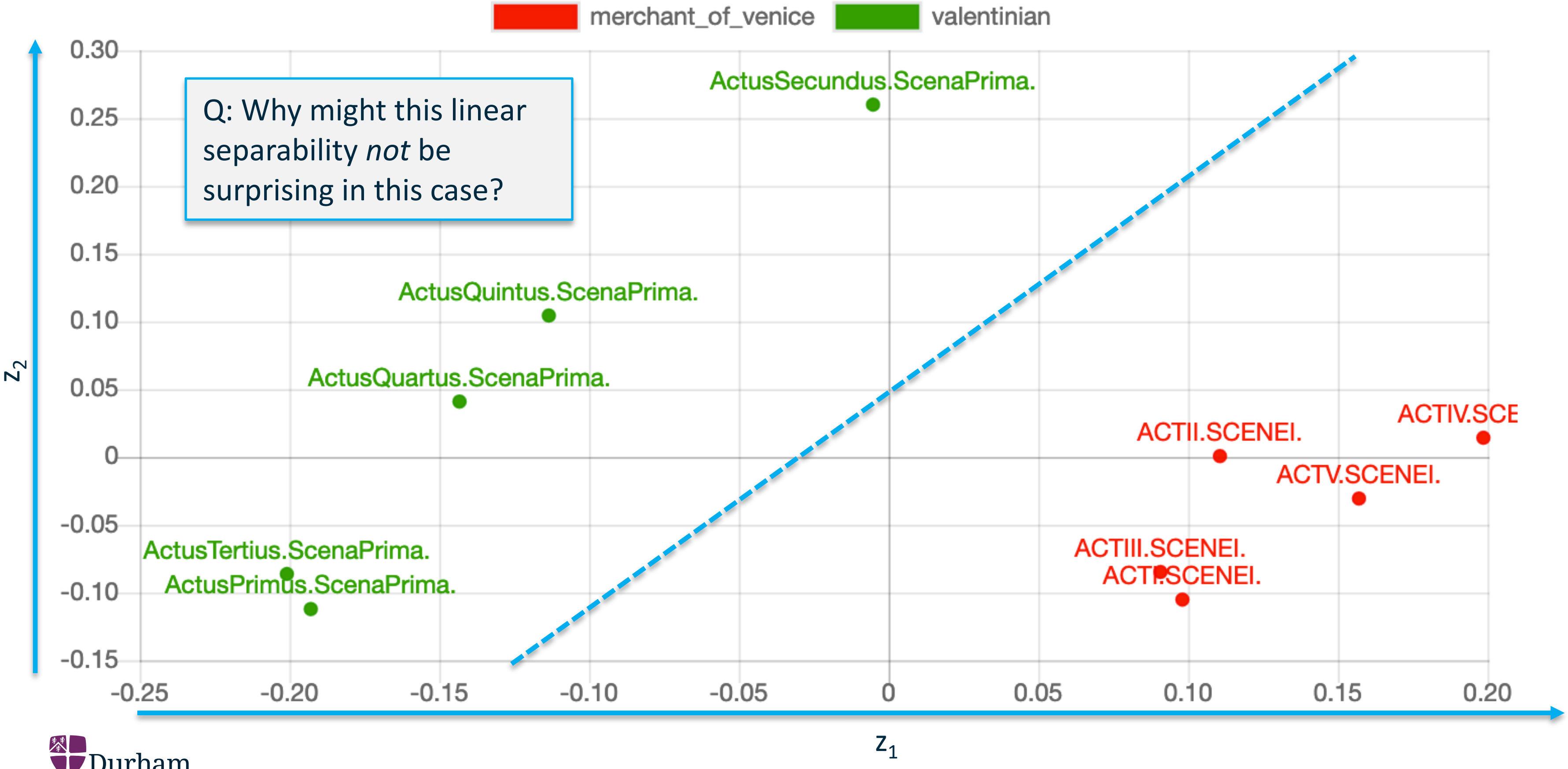
...

...

yours-	0	0	0.00249
yourself	0.00168	0	0.00249
Yourself	0	0.00116	0
youth	0.00503	0	0.00249
Youth	0	0	0
youthful	0	0	0
zeal	0	0	0

...

Authorship attribution – example (TF vectors, by act)



Q: Why might this linear separability *not* be surprising in this case?

Authorship attribution – example

Linear transformation z = Px, $z \in \mathbb{R}^N$, $x \in \mathbb{R}^N$

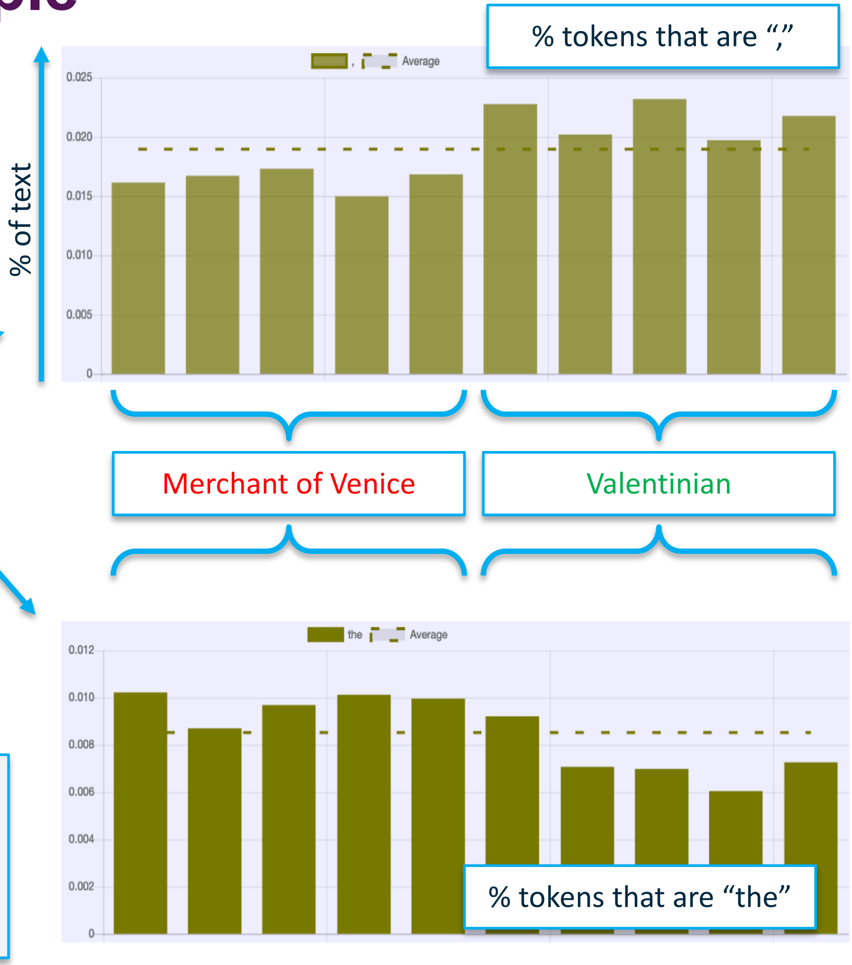
$$z_i = \sum_{j=1}^{N} P_{ij} x_j$$

In each axis in the projection, a weight is assigned to a vector component – in our example, this corresponds to a specific term

Sorted by $|P_{1j}|$

V_j	P_{1j}
,	-0.453
.	0.378
the	0.319
ye	-0.246
you	0.232
:	-0.191
;	0.167
…	…

z_1 (horizontal coordinate in last slide) is the sum of all these values multiplied by the TF of each corresponding term



% tokens that are “,”

Merchant of Venice

Valentinian

% tokens that are “the”

Durham University

Authorship attribution – example

Linear transformation $z = Px$, $z \in \mathbb{R}^N$, $x \in \mathbb{R}^N$

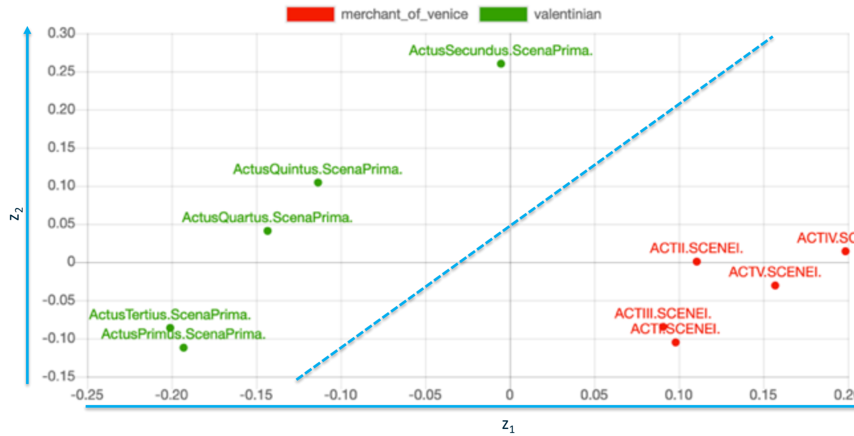
$$z_i = \sum_{j=1}^{N} P_{ij} x_j$$

In each axis in the projection, a weight is assigned to a vector component – in our example, this corresponds to a specific term

Sorted by $|P_{1j}|$

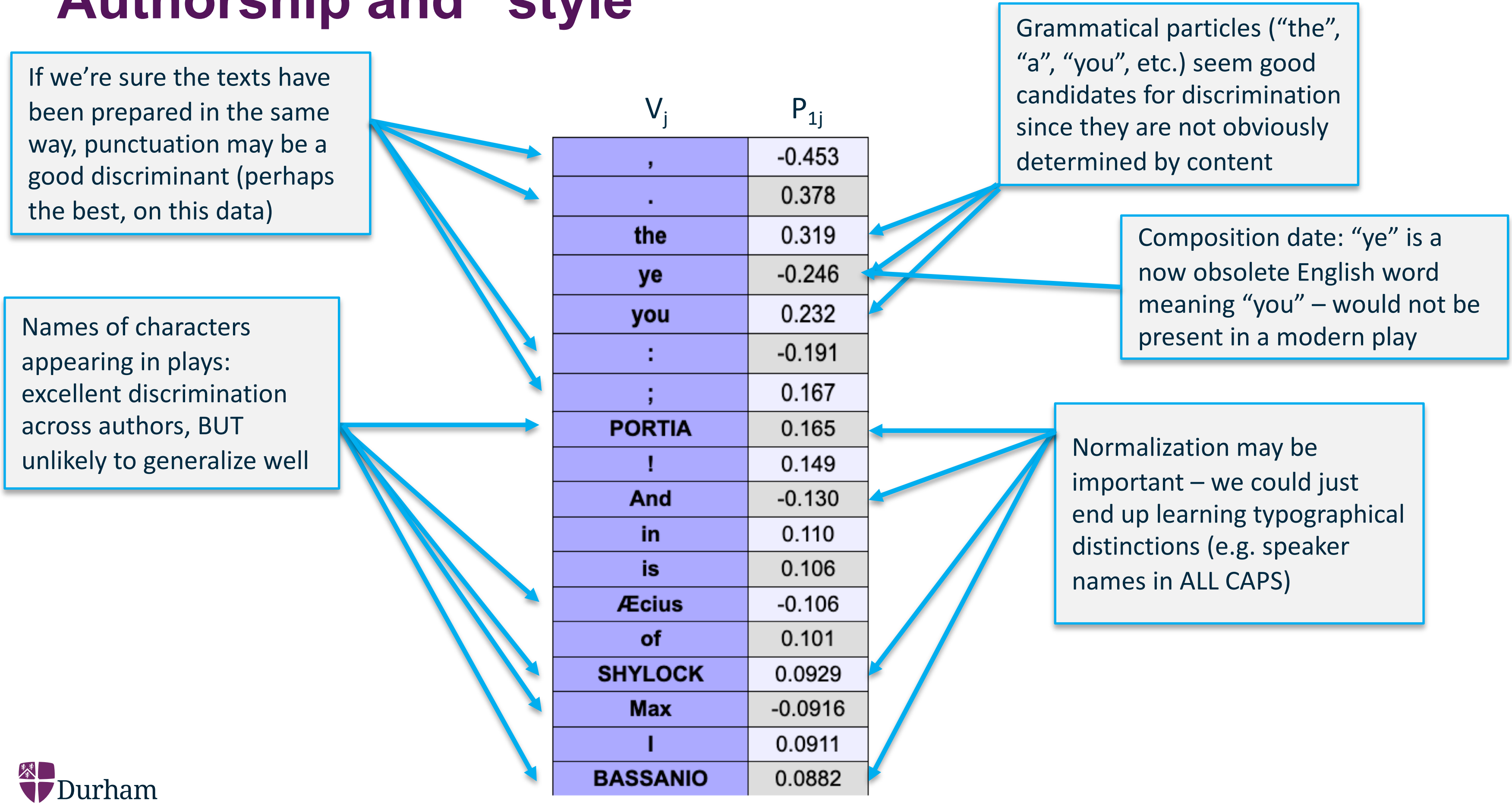
V_j	P_{1j}
,	-0.453
.	0.378
the	0.319
ye	-0.246
you	0.232
:	-0.191
;	0.167
…	…

z_1 (horizontal coordinate in last slide) is the sum of all these values multiplied by the TF of each corresponding term



I.e. if we write t_c for term frequency of token c, then:
$z_1 = -0.453t_{,} + 0.378t_{.} + 0.319t_{the} – 0.246t_{ye} + ….$

Authorship and “style”



If we’re sure the texts have been prepared in the same way, punctuation may be a good discriminant (perhaps the best, on this data)

Names of characters appearing in plays: excellent discrimination across authors, BUT unlikely to generalize well

V_j	P_{1j}
,	-0.453
.	0.378
the	0.319
ye	-0.246
you	0.232
:	-0.191
;	0.167
PORTIA	0.165
!	0.149
And	-0.130
in	0.110
is	0.106
Æcius	-0.106
of	0.101
SHYLOCK	0.0929
Max	-0.0916
I	0.0911
BASSANIO	0.0882

Grammatical particles (“the”, “a”, “you”, etc.) seem good candidates for discrimination since they are not obviously determined by content

Composition date: “ye” is a now obsolete English word meaning “you” – would not be present in a modern play

Normalization may be important – we could just end up learning typographical distinctions (e.g. speaker names in ALL CAPS)

Authorship and “style”

Pitfalls:

- Classifying correctly does not imply classification *by style*!
  - Could be by *content* or anything else correlated with features (inc noise)
  - Could be genre (e.g. poetry vs prose)
  - Models explainable with reference to e.g. *textual* features desirable
- Common difficulty: not enough independent samples from each author
  - E.g. can segment long works, but must ensure testing uses only *entirely* unseen texts (i.e. not unseen samples from a seen text)