Durham University

Computational approaches to textual similarity

Donald Sturgeon
Department of Computer Science
Durham University

Related materials and slides download: https://dsturgeon.net/cuhk2024

Overview

Types of similarity

- Sequential similarities – “text reuse”
- Similarities of vocabulary use

Simple approaches to identifying similarities

- N-grams
- Vector space model
  - Cosine similarity
  - Principal Component Analysis (PCA)

Related materials and slides download: https://dsturgeon.net/cuhk2024

Motivation

Similarities between texts can shed light on questions of textual history, authorship, etc.

- Tracing historical connections between written works through reuse of phrases/ideas
- Individual instances of reuse may be of interest to textual scholars
  - E.g. multiple “witnesses” to the same text, which may *disagree* on some of the details
- Similarities of both types have a long history of use in traditional textual scholarship

Why study these similarities *computationally*?

- Scale – a computer can exhaustively identify similarities of a particular type
- Subtlety – some similarities consist of statistical trends that are hard to see directly

Related materials and slides download: https://dsturgeon.net/cuhk2024

Text reuse

Global text reuse

Google Arts & Culture

Sign in

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

WIKIPEDIA

Search Wikipedia

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

FANDOM GAMES MOVIES TV VIDEO WIKIS

Information

Scientific Name Felis catus

Appearance

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.

Text reuse

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]

WIKIPEDIA Search Wikipedia

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

Text reuse: examples

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？」

Text reuse: examples

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

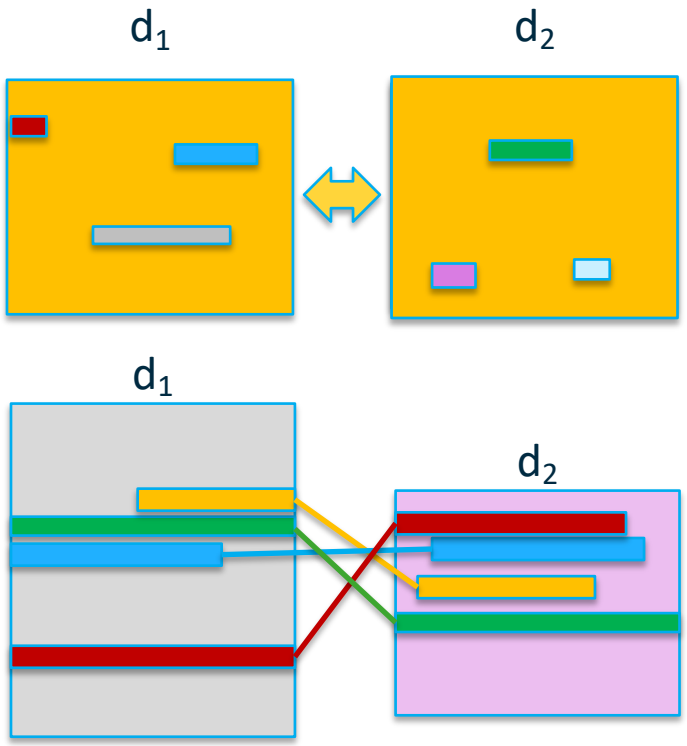
Global vs local reuse

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



Durham University

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

Identifying text reuse

N-grams

- Useful for both global and local similarity
- Can identify extended local similarities (e.g. quotation, copying, etc.)

Document vectors

- Simple and fast to calculate
- Ignores word order – so-called “bag of words” approach
- More useful for *global* similarities

Other approaches

- Can be tailored to identifying particular types of similarity
- Often use precision/recall to evaluate the level of success at identifying

Tokenization

- When comparing text, often the most intuitive units to compare are *words*

Using *words* as tokens

S_1 = `anthropological explorations and examples of creatures considered pests`

S_2 = `ants rats and fleas are examples of creatures considered pests`

Using *characters* as tokens

S_1 = `anthropological explorations and examples of creatures considered pests`

S_2 = `ants rats and fleas are examples of creatures considered pests`

Tokenization

S_1 = 保持共產黨員先進性教育活動

T_1 =	保持	共產黨員	先進性	教育	活動
	uphold	Communist party members	advanced nature	education	campaign
T_2 =	保持	共產黨員	先進	性教育	活動
	uphold	Communist party members	advanced	sex education	campaign

- Both syntactically valid tokenizations
- Different meanings

- Tokenization is not trivial for all languages (e.g. Chinese)

N-grams

- “n” is simply some fixed integer, e.g. n=1；n=2；n=3；…
- An “n-gram” is a sequence of *n* tokens* appearing in sequence
  - 1-gram = one token (i.e. 1 character, or 1 word)
  - 2-gram = two tokens
  - 3-gram = three tokens
  - …
- * “Tokens” are the smallest unit of text we will work with. These could be words; for classical Chinese these will often be characters.

Example: n-grams, n=1, characters as tokens

天命之謂性，率性之謂道，修道之謂教。

1-grams: which unique characters appear, and how many times each?

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

Example: n-grams, n=2, characters as tokens

天命之謂性，率性之謂道，修道之謂教。

2-grams: which unique *pairs* of characters appear, and how many times each?

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

Example: n-grams, n=3, characters as tokens

天命之謂性，率性之謂道，修道之謂教。

3-grams: which unique *triples* of characters appear, and how many times each?

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

N-grams capture some aspects of word use

《史記》		《荀子》		《詩經》	
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

Normalization

天下	1218
諸侯	964
將軍	865
以為	846
於是	843

《史記》

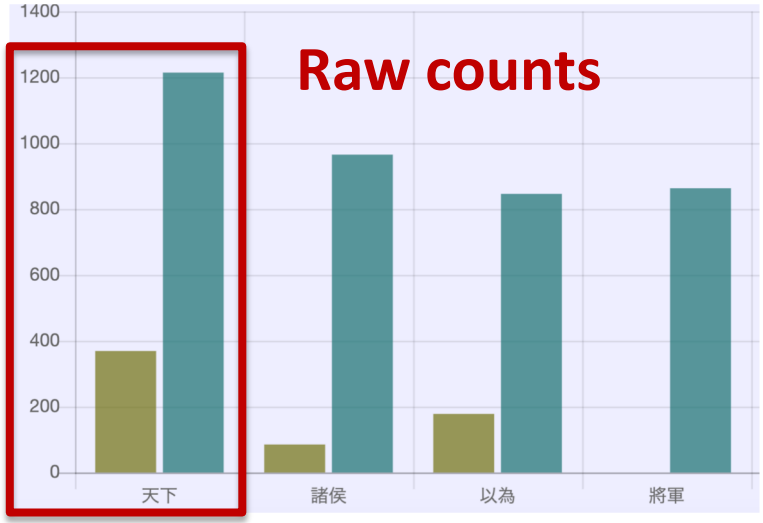
>500,000 characters

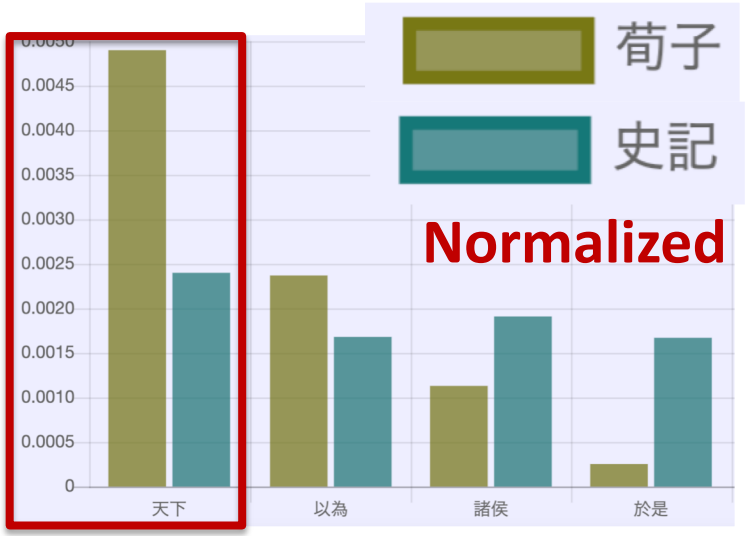
而不	382
天下	372
人之	300
君子	299
也故	240

《荀子》

75,000 characters

Which uses “天下” more?





Durham University

Using n-grams to identify text reuse

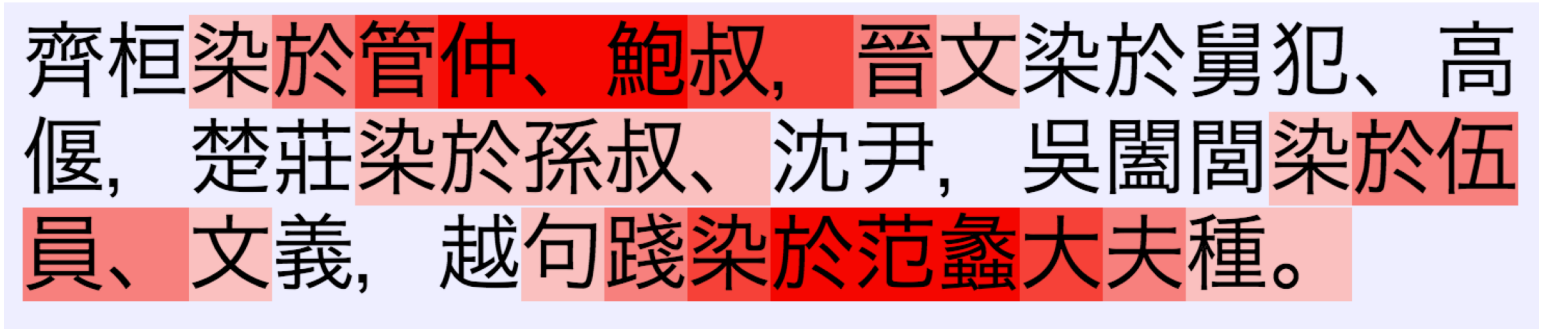
- Example: n=4
- 齊桓染於管仲、鮑叔，晉文染於舅犯、高偃，楚莊染於孫叔、沈尹，吳闔閭染於伍員、文義，越句踐染於范蠡大夫種。

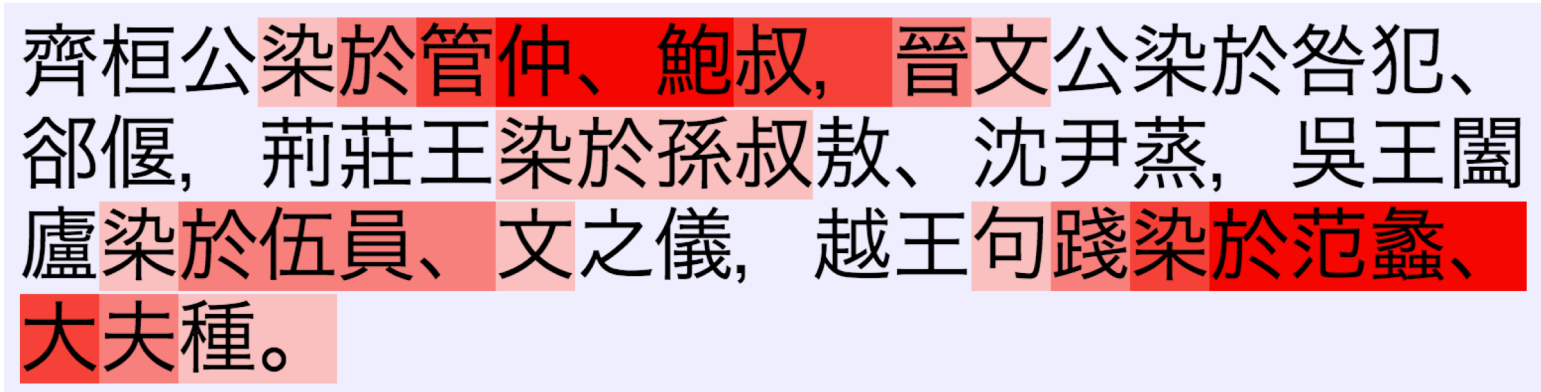
- 齊桓公染於管仲、鮑叔，晉文公染於咎犯、郤偃，荊莊王染於孫叔敖、沈尹蒸，吳王闔廬染於伍員、文之儀，越王句踐染於范蠡、大夫種，

Durham University

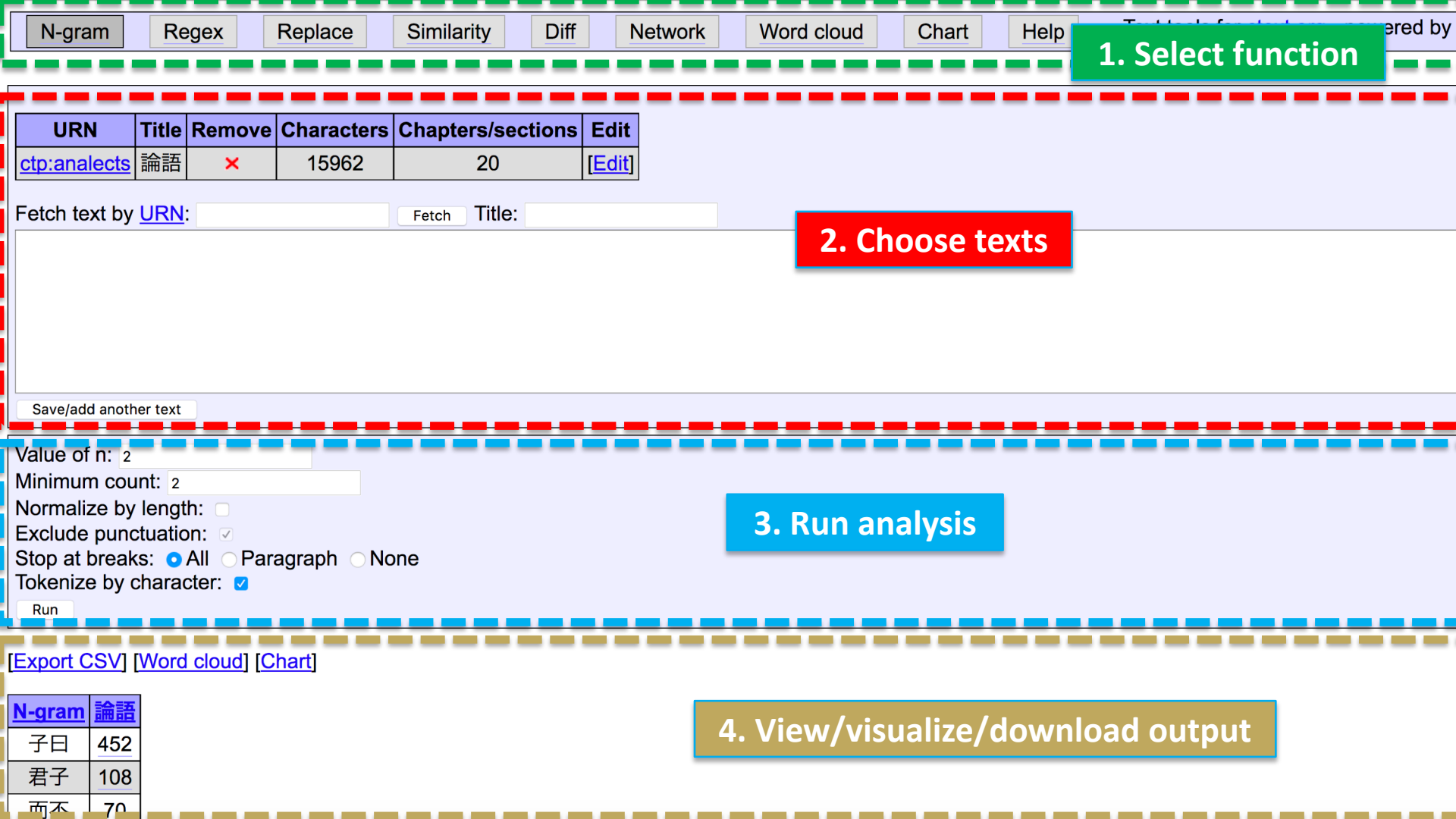
Using n-grams to identify text reuse

- Example: n=4





Durham University


N-gram
Regex
Replace
Similarity
Diff
Network
Word cloud
Chart
Help
1. Select function
URN
Title
Remove
Characters
Chapters/sections
Edit
ctp:analects
論語
15962
20
[Edit]
Fetch text by URN:
Fetch
Title:
2. Choose texts
Save/add another text
Value of n: 2
Minimum count: 2
Normalize by length:
Exclude punctuation:
Stop at breaks: All Paragraph None
Tokenize by character:
Run
3. Run analysis
[Export CSV] [Word cloud] [Chart]
N-gram
論語
子曰
452
君子
108
4. View/visualize/download output

Hands-on examples

1. Text reuse in the Analects
   - Open https://text.tools/ctext
   - Download analects.zip
   - Drag analects.zip onto the Text Tools page
   - Click “Similarity” to select the n-gram similarity function
   - Click “Run” to run the comparison with the default settings
2. Experiment with interaction, visualizations, and other texts

Hands-on examples

3. English example with tokenization
   - Open https://text.tools/ctext (or remove loaded texts)
   - Download alice.zip
   - Drag alice.zip onto the Text Tools page
   - Click “Similarity” to select the n-gram similarity function
   - Unselect “Tokenize by character”
   - Click “Run” to run the comparison with the default settings

Similarity of vocabulary

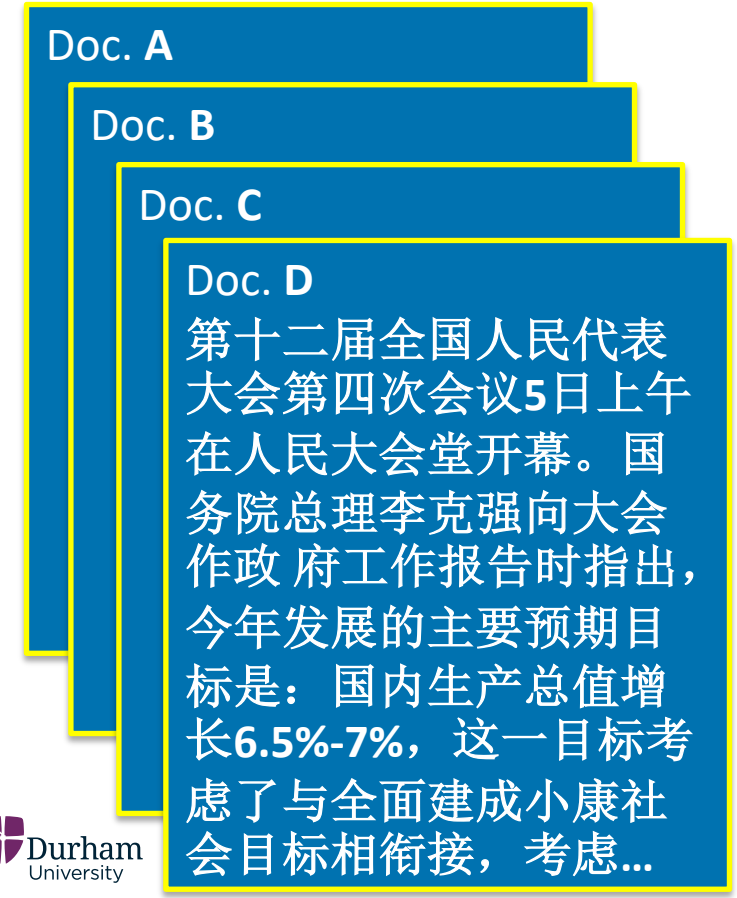
Exploratory approaches:

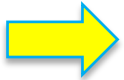
- Compare all vocabulary use among a set of texts
- Can give indication of “unusual” word usage within a corpus

Targeted approach:

- Explore how some chosen set of words vary across a corpus
- Useful where there is some hypothesis about word usage
- Can be used to investigate aspects of authorial style

Document vectors





Doc. A	Doc. B	Doc. C	Doc. 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

We will calculate similarity between vectors, corresponding to similarity between their documents

Durham University

Document vectors

To create one vector to represent each document:

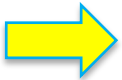
1. First decide on one “vocabulary” to use for all documents
    - Method 1: choose all unique words that appear in any document
    - Method 2: choose a hand-chosen list of words we wish to consider
2. For each document:
    - For every word (token) in our vocabulary:
        - Write down how many times that token occurs in this document

This approach gives us “Term frequency vectors”

3. Document vectors: example

Doc. A

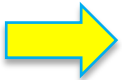
这句话只是一个例子，
只是想说明向量是什么。



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

Doc. B

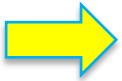
这也是一个例句。



这 / 也 / 是 / 一 / 个 / 例句 / 。

Doc. C

举个例子吧。



举 / 个 / 例子 / 吧 / 。

Document vectors: example

Doc. A

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

Doc. B

这 / 也 / 是 / 一 / 个 / 例句 / 。

Doc. C

举 / 个 / 例子 / 吧 / 。

Vocab	Doc. A	Doc. B	Doc. 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	1
吧	0	0	1

Similarity of document vectors

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

Frequently used metric: cosine similarity

Vocab	Doc. A	Doc. B	Doc. 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

Doc. A

这 / 一 / 句 / 话 / 只 / 是 / 一 / 个 / 例子

Doc. B

这 / 是 / 一 / 个 / 例子

Doc. C

另 / 一 / 个 / 话题

Similarity of document vectors

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

Frequently used metric: cosine similarity

Vocab	Doc. A		Doc. 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

Doc. A

这 / 一 / 句 / 话 / 只 / 是 / 一 / 个 / 例子

Doc. B

这 / 是 / 一 / 个 / 例子

Doc. C

另 / 一 / 个 / 话题

Similarity of document vectors

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

Frequently used metric: cosine similarity

Vocab	Doc. A			Doc. 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

Doc. A

这 / 一 / 句 / 话 / 只 / 是 / 一 / 个 / 例子

Doc. B

这 / 是 / 一 / 个 / 例子

Doc. C

另 / 一 / 个 / 话题

Similarity of document vectors

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

Frequently used metric: cosine similarity

Vocab	Doc. A	Doc. B	Doc. 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

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

$= 0.809$

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

$= 0.452$

B is more similar to A than C is

Text A

这 / 一 / 句 / 话 / 只 / 是 / 一 / 个 / 例子

Text B

这 / 是 / 一 / 个 / 例子

Text C

另 / 一 / 个 / 话题

Durham University

Cosine similarity and vector space

Document 1: cat cat cat cat

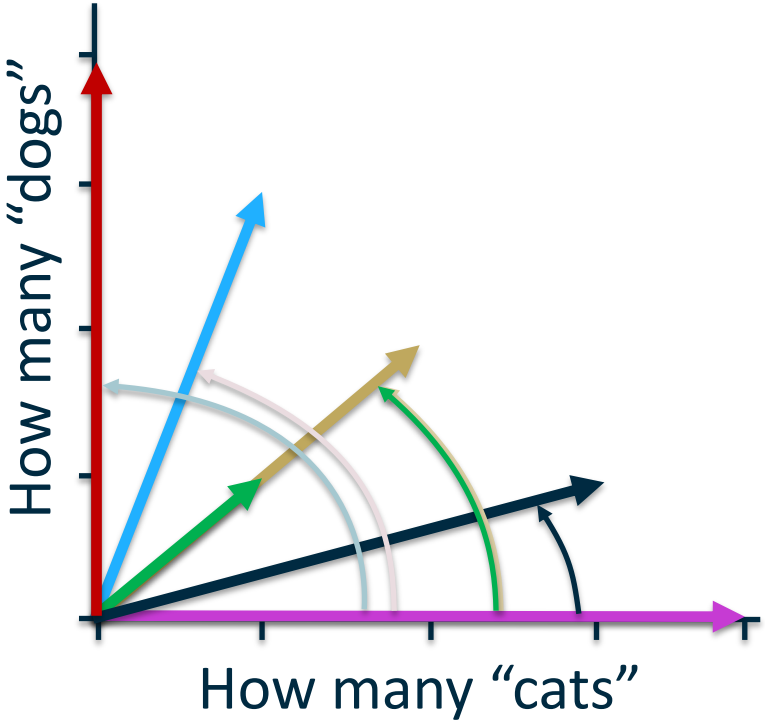
Document 2: cat dog

Document 3: dog dog cat dog

Document 4: dog dog cat cat

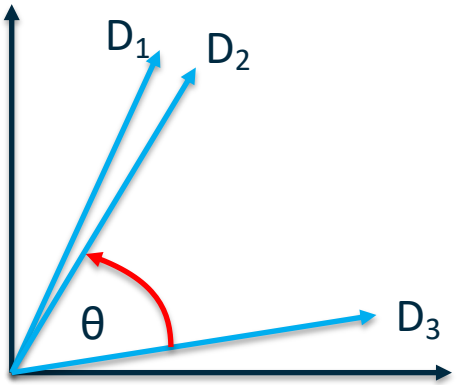
Document 5: cat cat cat dog

Document 6: dog dog dog dog



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

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

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

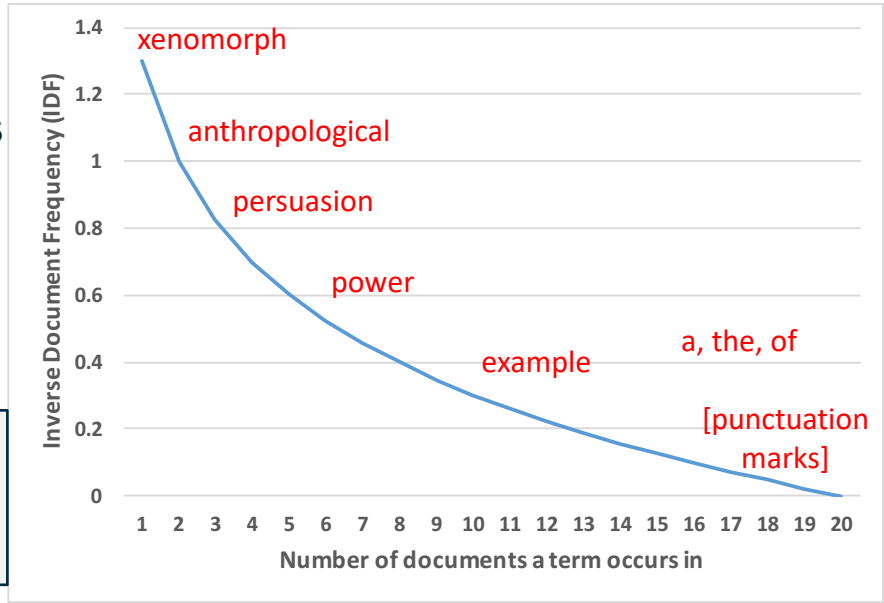
Easy approaches:

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

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

Document vectors

Usually very sparse

- Most components of most vectors are 0
  - Why? Most documents do not contain most of the vocabulary!

Useful where “bag of words” assumption is reasonable

- Order and context less important than frequent mentions of terms
- vs e.g. text reuse => vectors can only be used to detect reused *vocabulary*

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

Cosine similarity – example

Similarity matrix:

Each cell colored by $sim_{cosine}(D_i, D_j)$

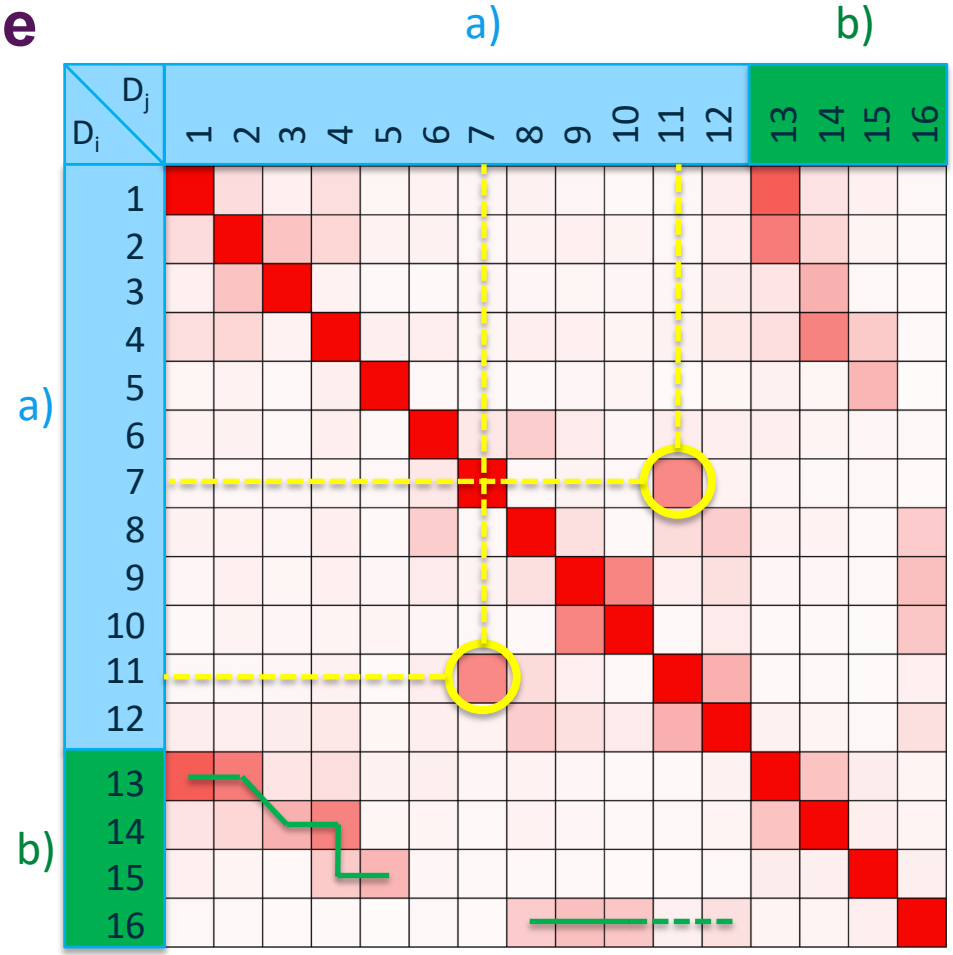
Solid white: 0 solid red: 1

Blue & green two document groups

a) Alice in Wonderland

b) Alice's Adventures Underground

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

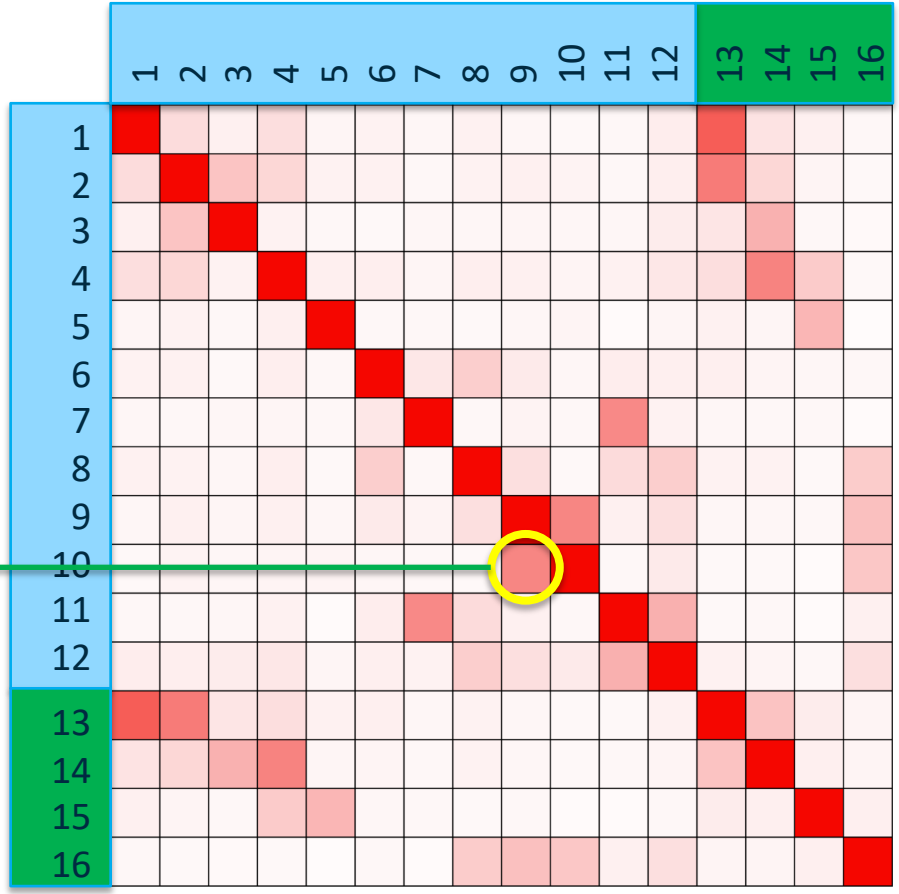


Cosine similarity – example
Interpretability

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

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



Durham University

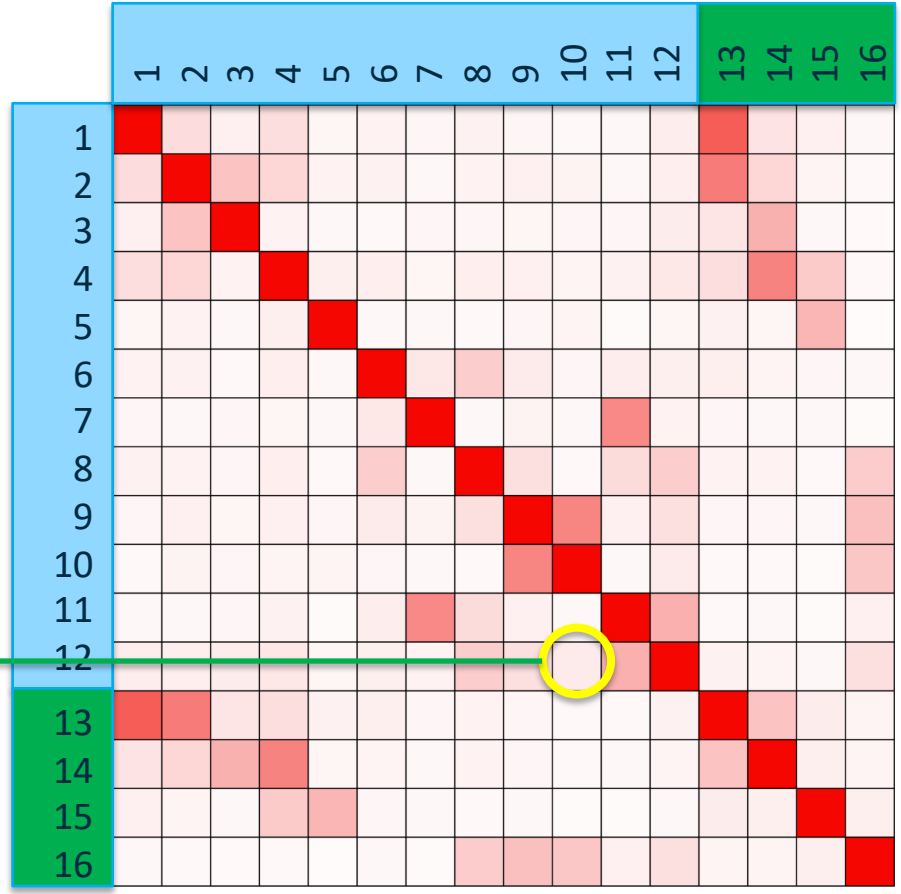
Cosine similarity – example
Interpretability

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

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!



Durham University

Hands-on examples

- Chinese examples:
  - Load the texts
  - “Vectors” -> “Run”
- Alice in Wonderland:
  - Load the texts
  - “Transform” -> “Register” -> “English Tokenizer (Moses)” -> “Apply to all”
  - “Vectors” -> (uncheck “Tokenize by character) -> “Run”

Principal Component Analysis (PCA)

- In modeling texts, we often use *vectors* to represent documents
  - We used these to represent text digitally
  - We used these to identify textual similarity
  - 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
…	…	…

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 we only used the first 3 words:

doc1=[1,2,0]

doc2=[0,1,1]

If we only used the first 2 words:

doc1=[1,2]

doc2=[0,1]

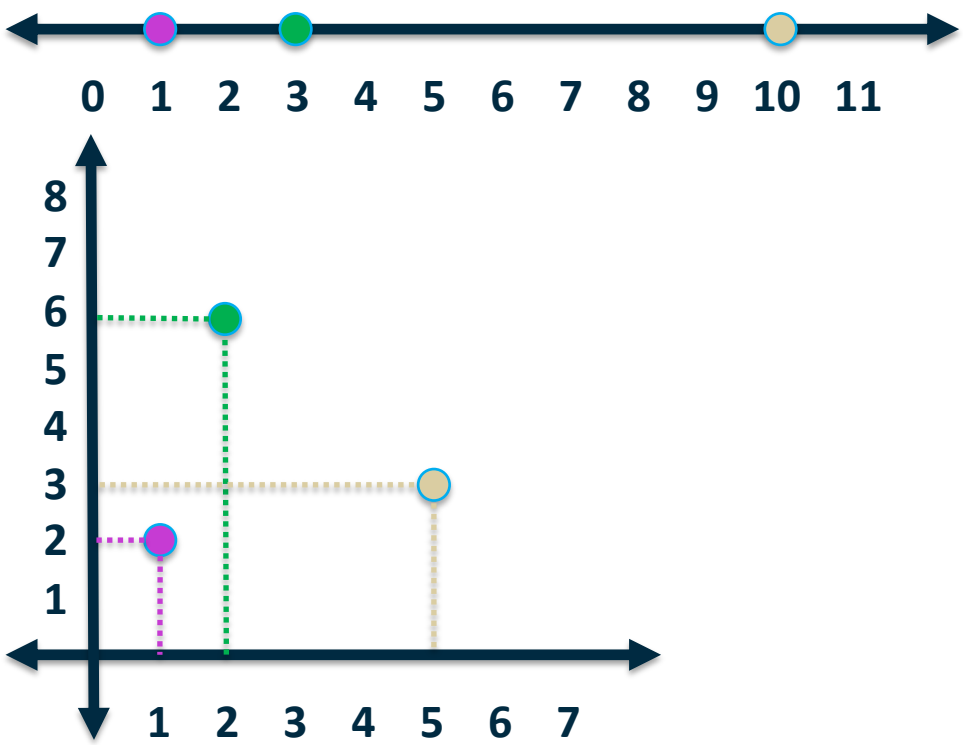
Dimensions and visualization

1 Dimension: x-coordinate

Example: [1]
[3]
[10]

2 Dimensions: x and y

Example: [1, 2]
[2, 6]
[5, 3]

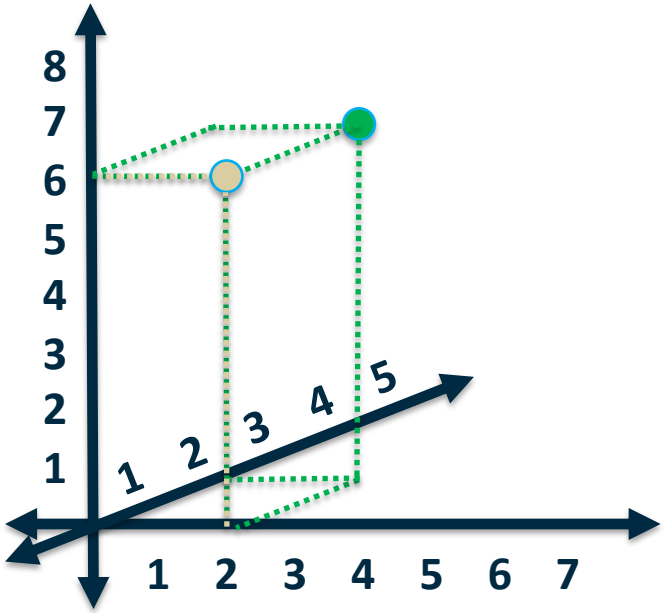


Dimensions and visualization

3 Dimensions: x, y, and z

Example: [2, 6, 0]

[2, 6, 2]



Computer screens are 2-dimensional. This example is a projection from 3 to 2 dimensions

Dimensions and visualization

4 Dimensions: x, y, z, a

Example: **[1, 2, 5, 1]**

[2, 6, 1, 9]

[5, 6, 3, 8]

PCA is a way of projecting data from high dimensions into lower dimensions

- It is calculated from a set of vectors
- It results in a *linear* projection
- It preserves as much variance as possible between data points

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

PCA and stylometry

PCA can be used to visualize patterns in how vocabulary is used

- Hypothesis of stylometry: authors subconsciously use different function words more or less often than other authors.
  - E.g. the, a, of, and, is, are, in, …

General idea:

- If all vectors from one author are linearly separable from those of another author, this indicates that their use of the chosen vocab is clearly distinct
- We can predict which of the two authors an unknown work is written by by comparing the vectors to the known samples

Example – Wizard of Oz

Two authors individually wrote a number of books in this series

- Frank Baum, and Ruth Plumly Thompson
- “Wizard of Oz”, and a large number of sequels
  - All belong to the same genre (fantasy novels)
  - Different characters and content
- One particular book in the series might have been written by either author

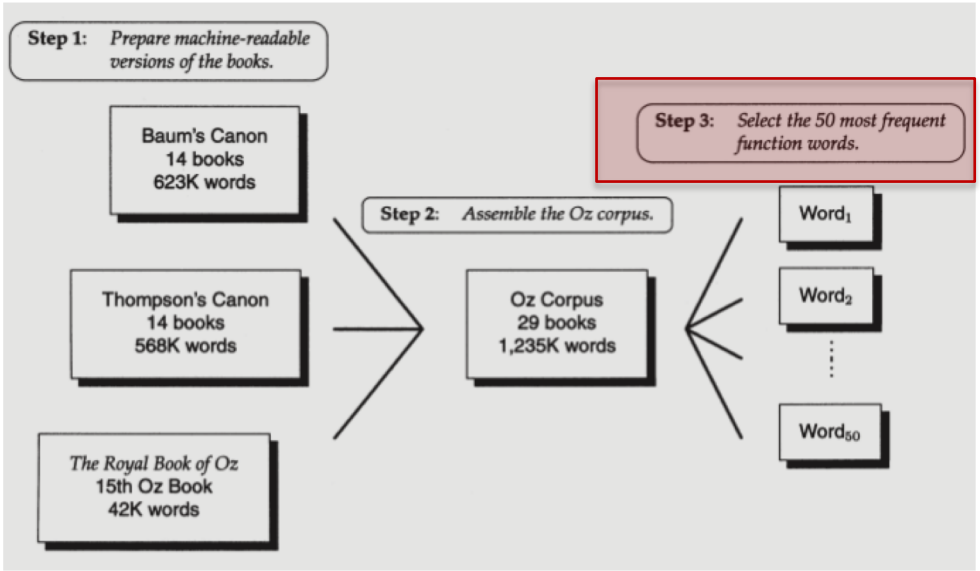
[Similar in principle to another well-known example, the “Federalist Papers”]

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?



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

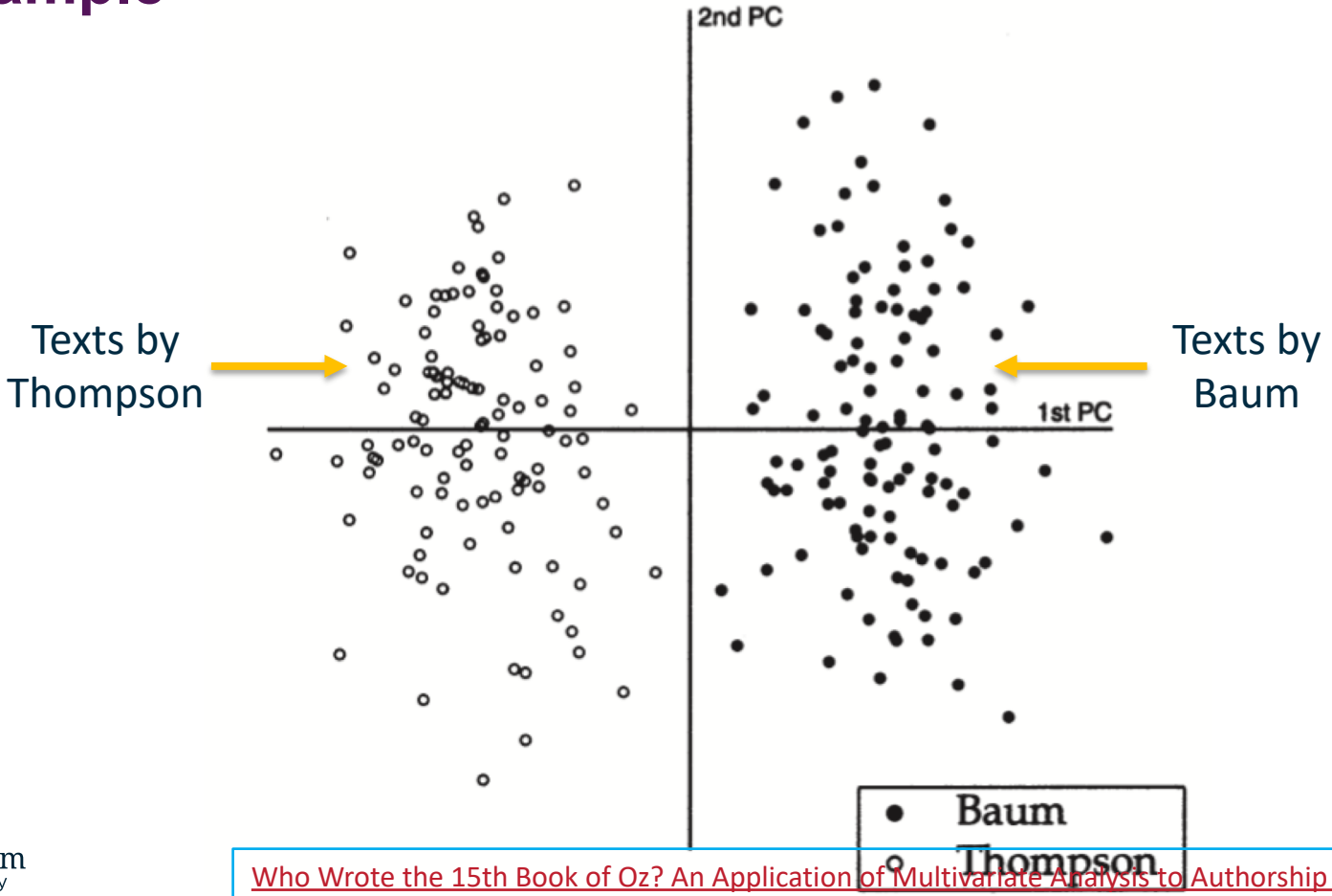
Example – Wizard of Oz

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 –



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

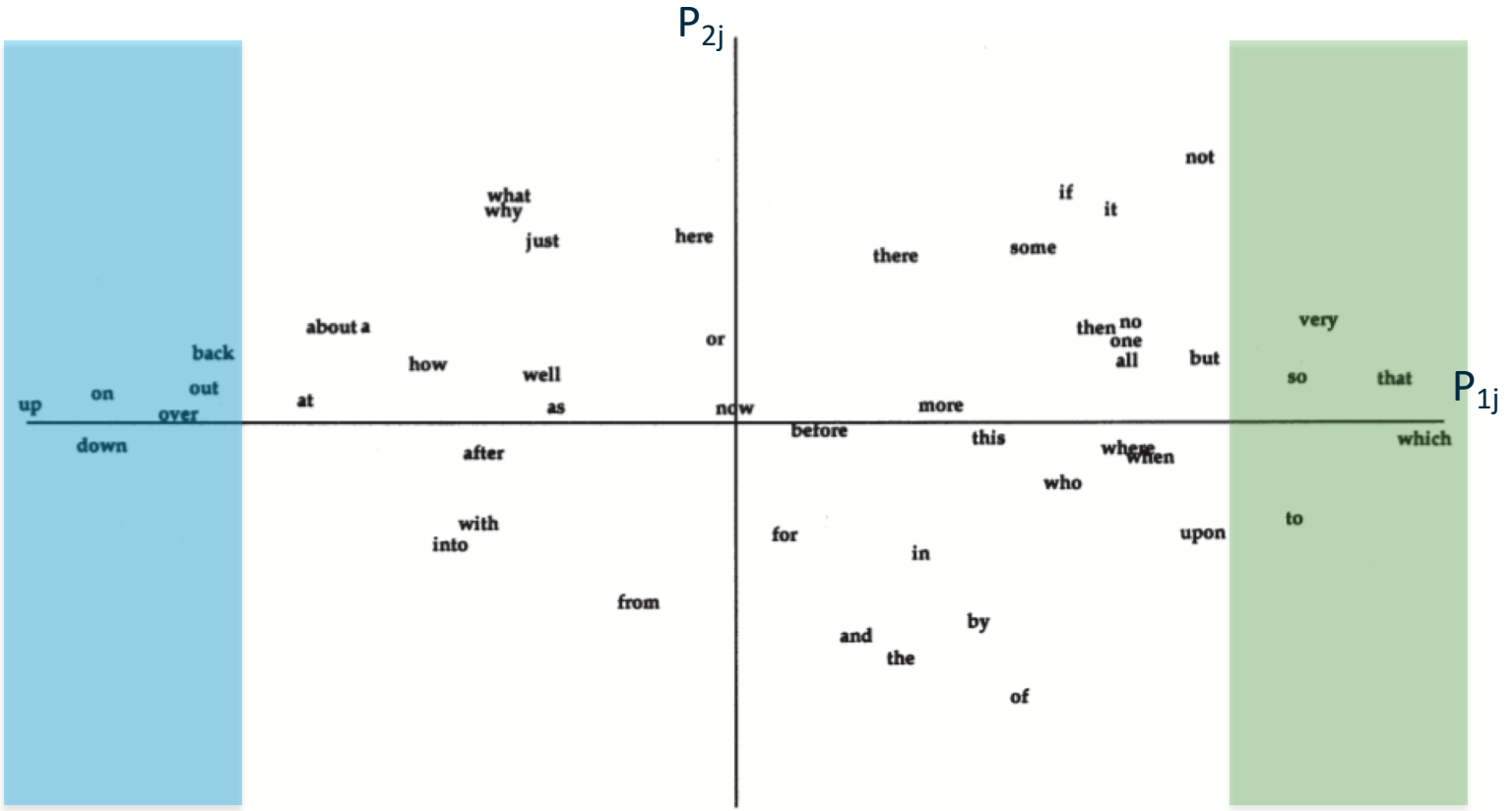


Figure 6. Component loadings.

Overrepresented in Thompson

Overrepresented in Baum

Durham University

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 .

Durham University

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

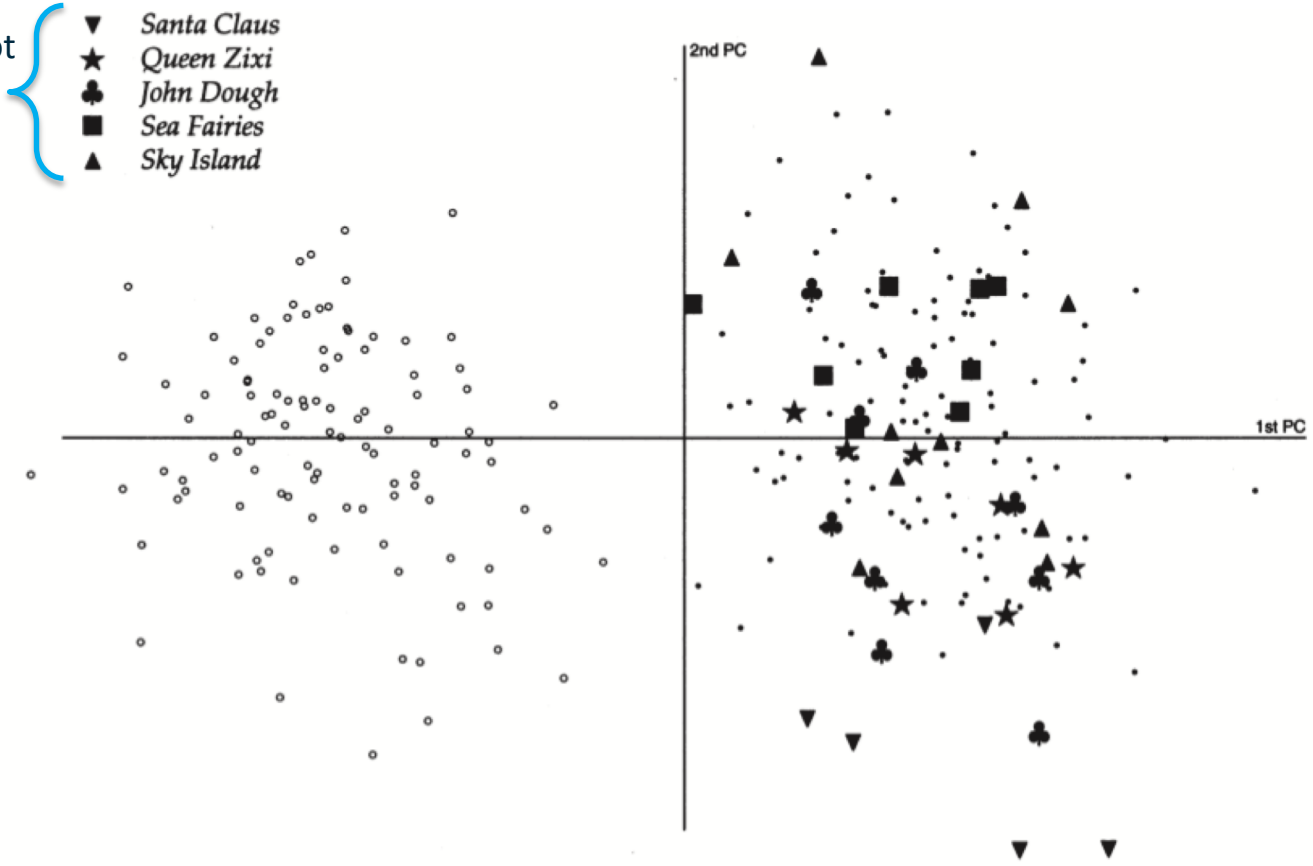


Figure 7. Baum's non-canonical works.

Durham University

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

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.

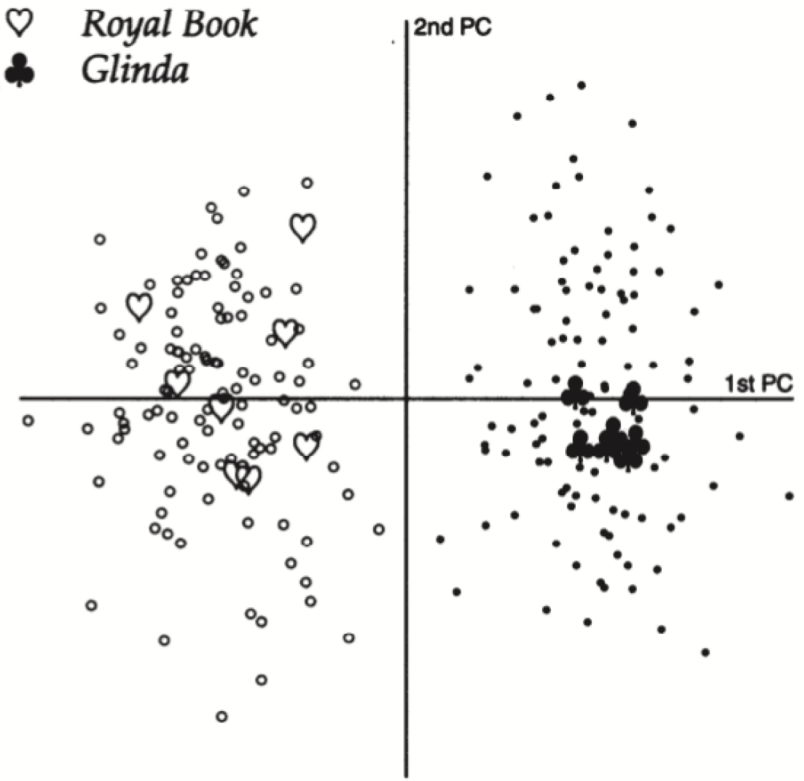
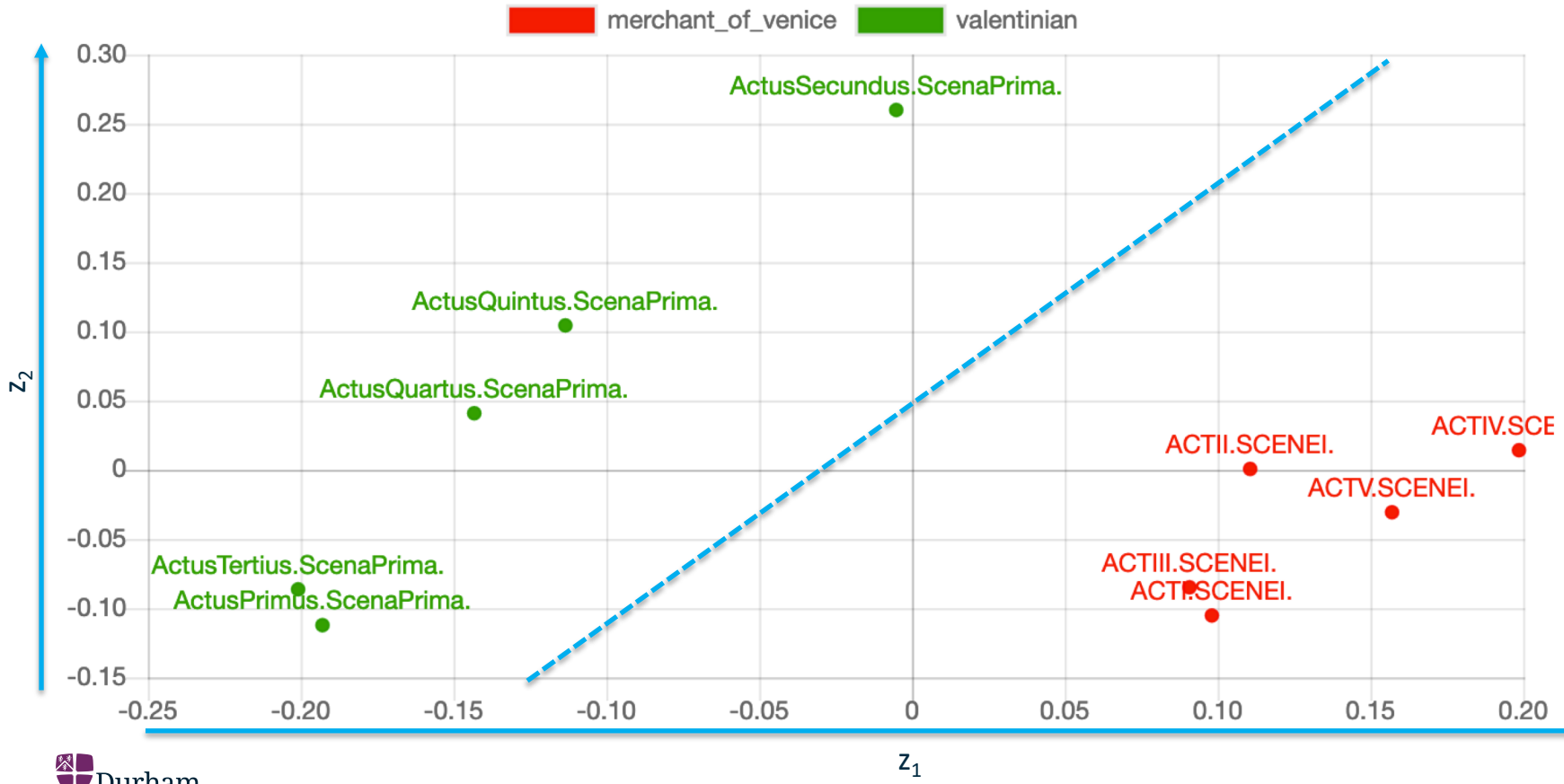


Figure 10. *The Royal Book of Oz* (1921).

Authorship attribution – example (TF vectors, by act)



Durham University

Authorship attribution – example

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

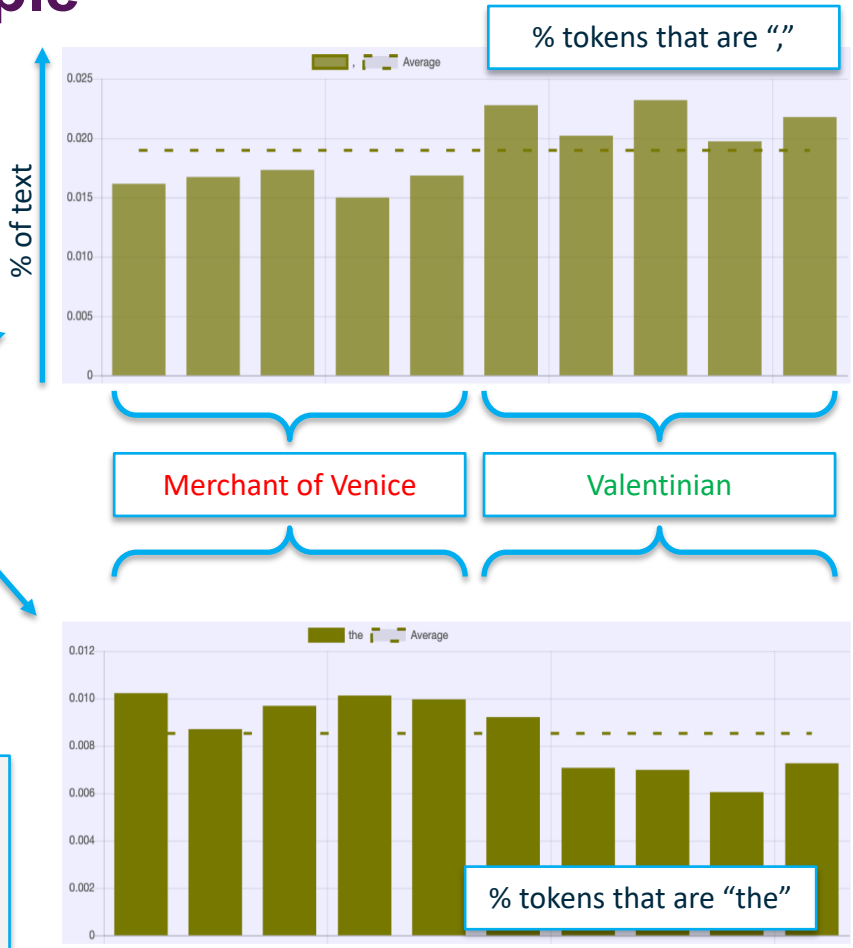
$$z_i = \sum_{j=1}^{N} P_{ij} x_j$$

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



Durham University

Authorship attribution – example

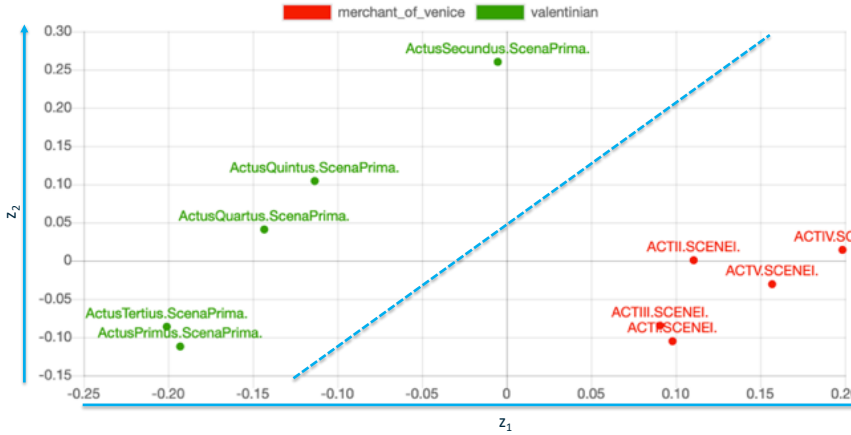
Linear transformation $z = Px$, $z \in \mathbb{R}^N$, $x \in \mathbb{R}^N$

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

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
...	...



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

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

Durham University

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)

Hands-on examples

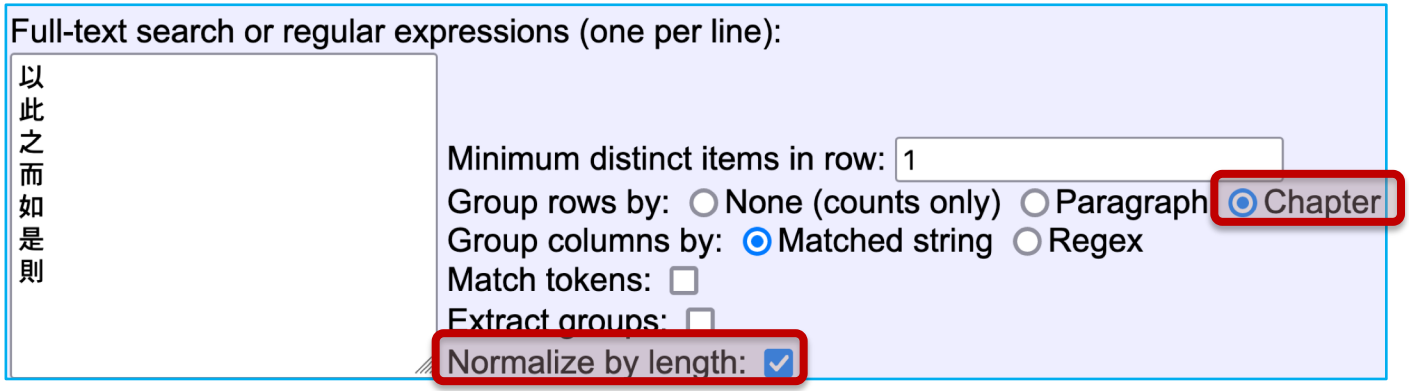
- Wizard of Oz
  - Load the texts
  - “Transform” -> “Register” -> “English Tokenizer (Moses)” -> “Apply to all”
  - “Lowercase English text” -> “Apply to all”
  - “Regex” -> paste list of words -> Group rows by chapter; Normalize by length; Match tokens -> “Run”
  - “Summary” -> “Create vectors” -> “Run PCA”

Hands-on: suggested experiments

Technique	Texts	Steps to reproduce
N-gram	墨子+莊子+荀子	Similarity => n=7 => Run => Try clicking highlighted parts; Click “Similarity matrix” => “Toggle values”; “Chapter summary” => “Create graph” => “Draw”; Double-click graph edges to jump to text
Cosine similarity	墨子	Vectors => Run => Toggle values => Click highlighted sections; Click specific term to see distribution in corpus; Click on the chart to see specific textual matches; Click “Vectors” tab to return to the original view

Hands-on: suggested experiments

Technique	Texts	Steps to reproduce
PCA	明代小说	Regex => In “Full-text search” input search terms, one per line (see below); choose “Chapter” and “Normalize by length”, then click “Run” Click “Summary” => “Create vectors” => “Run PCA”



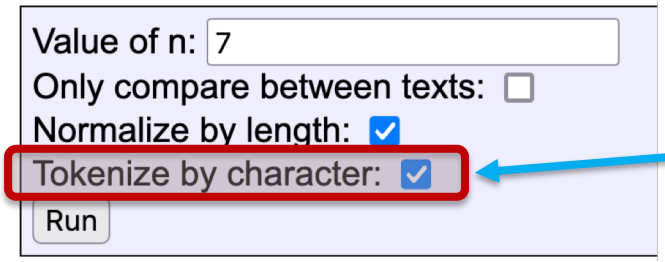
Durham University

Hands-on: suggested experiments

Technique	Texts	Steps to reproduce
PCA + TF-IDF (all vocab)	明代小说	Vectors => Run => Run PCA This will use all vocabulary in all texts

Hands-on: suggested experiments

Technique	Texts	Steps to reproduce
Tokenization	English / modern Chinese	Transform => “Register” => [choose type] => Run / Apply to all



N.B. When using tokenized texts, make sure to uncheck “Tokenize by character”