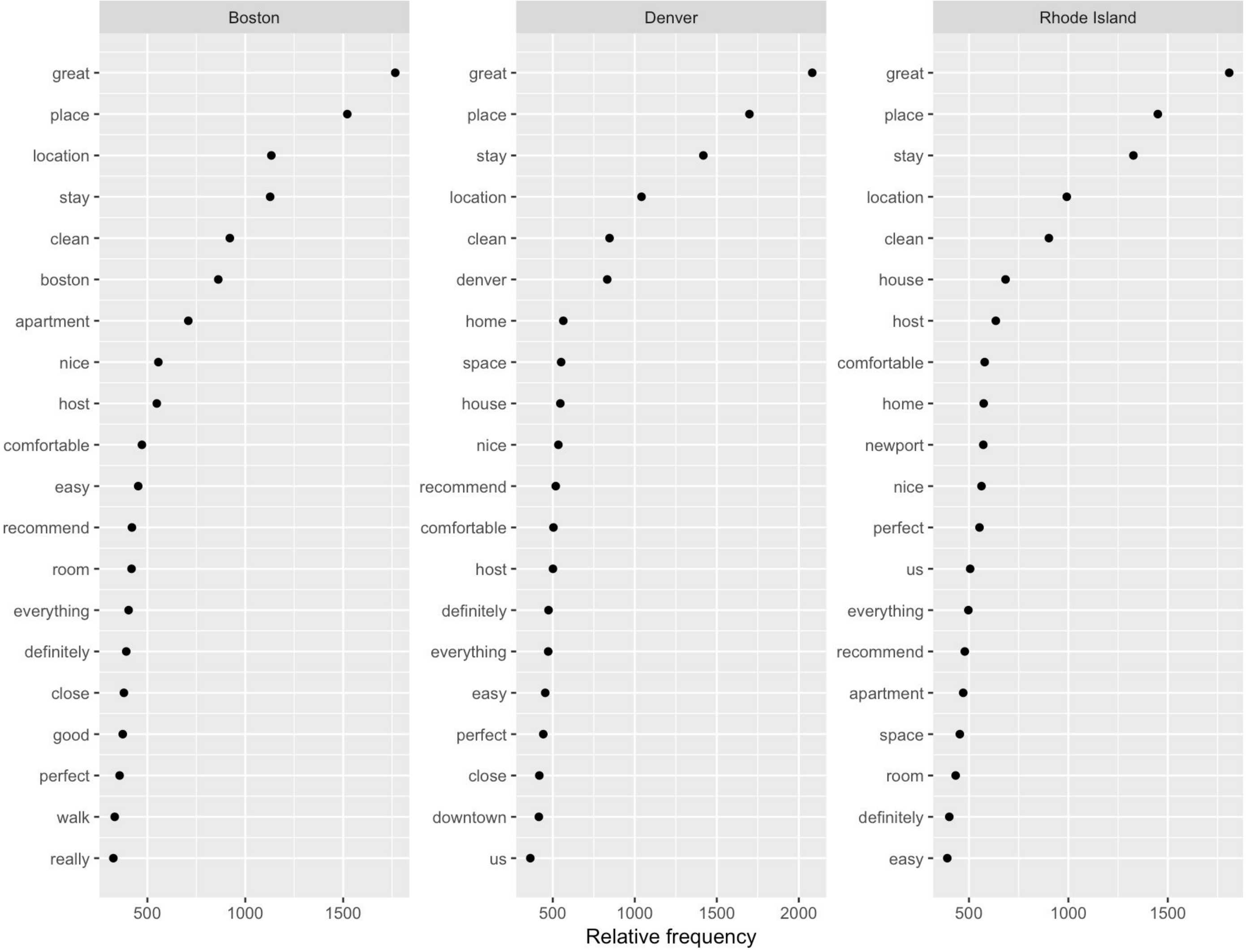
Text mining

Weiai Xu (Wayne), PhD
Assistant Professor
Department of Communication, UMass-Amherst
Email: weiaixu@umass.edu
curiositybits.cc

Things we will create

Frequency plot

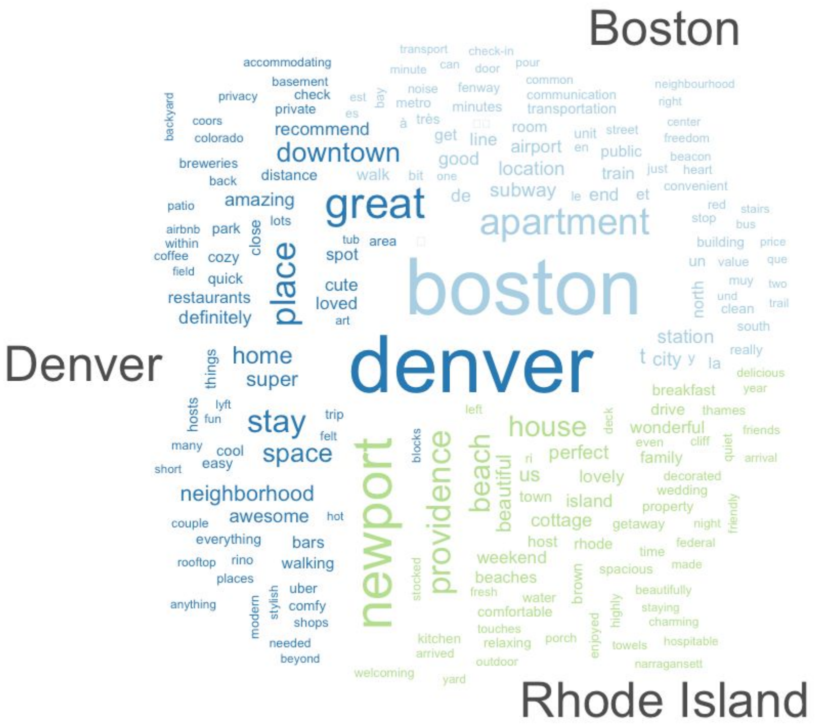
Based on the public Airbnb data: http://insideairbnb.com/get-the-data.html



Things we will create

Word cloud

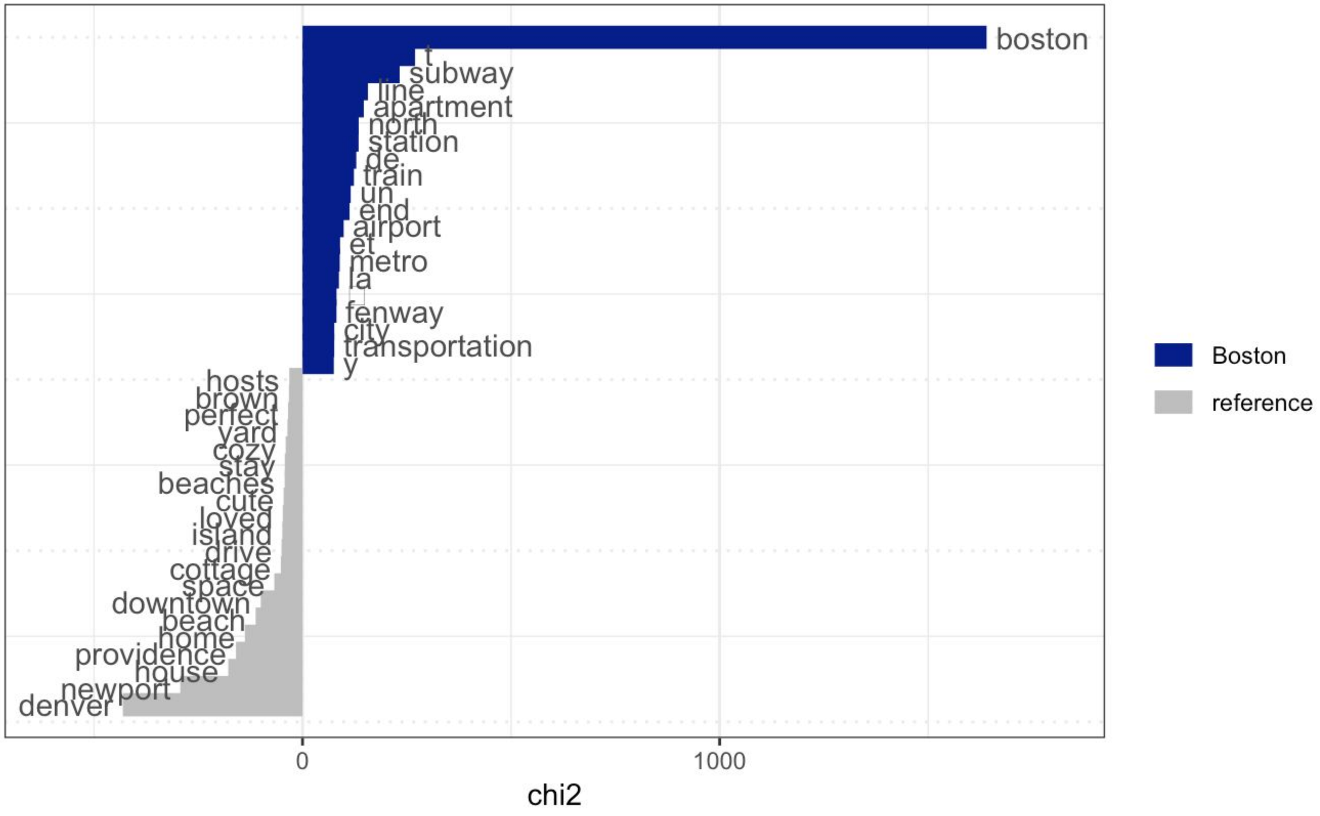
Based on the public Airbnb data: http://insideairbnb.com/get-the-data.html



Things we will create

Keyness

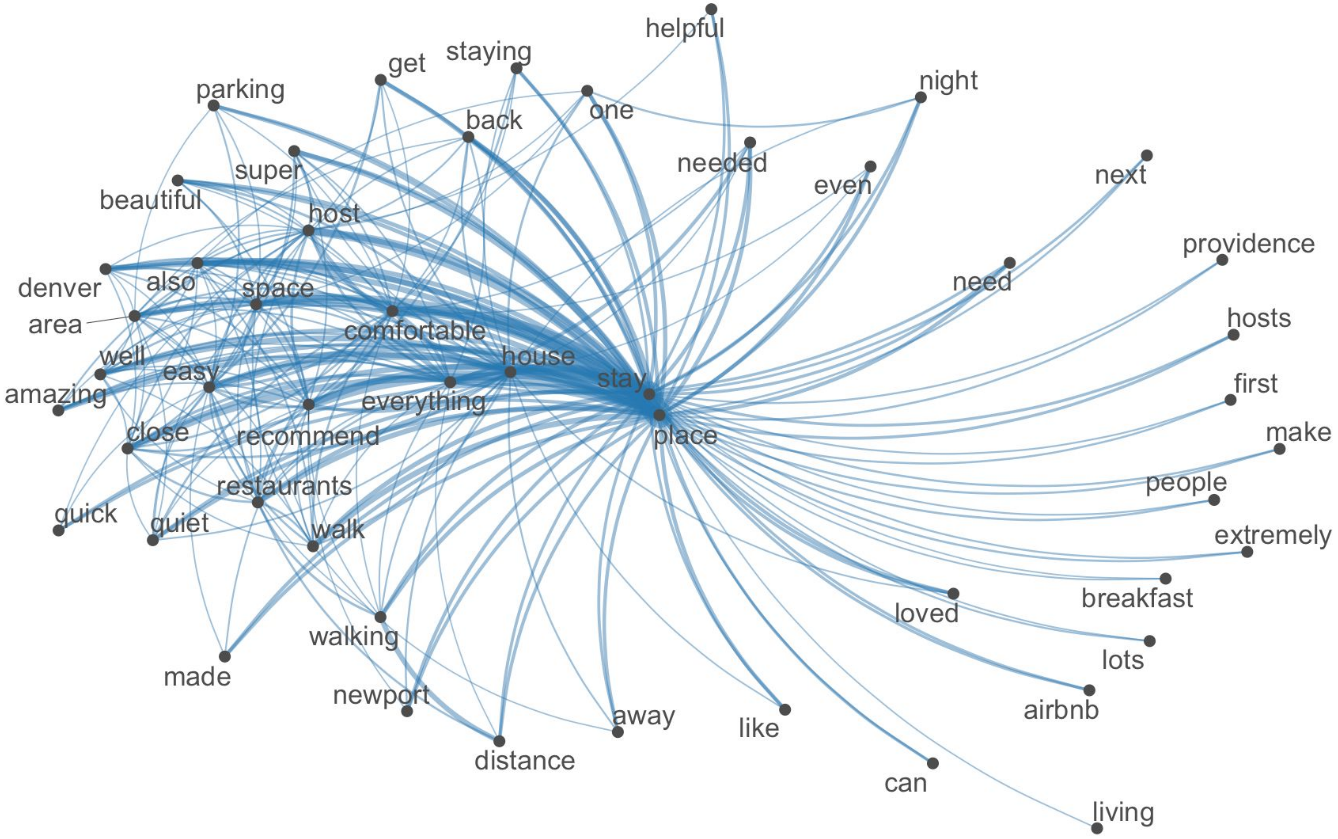
Based on the public Airbnb data: http://insideairbnb.com/get-the-data.html



Things we will create

Semantic network

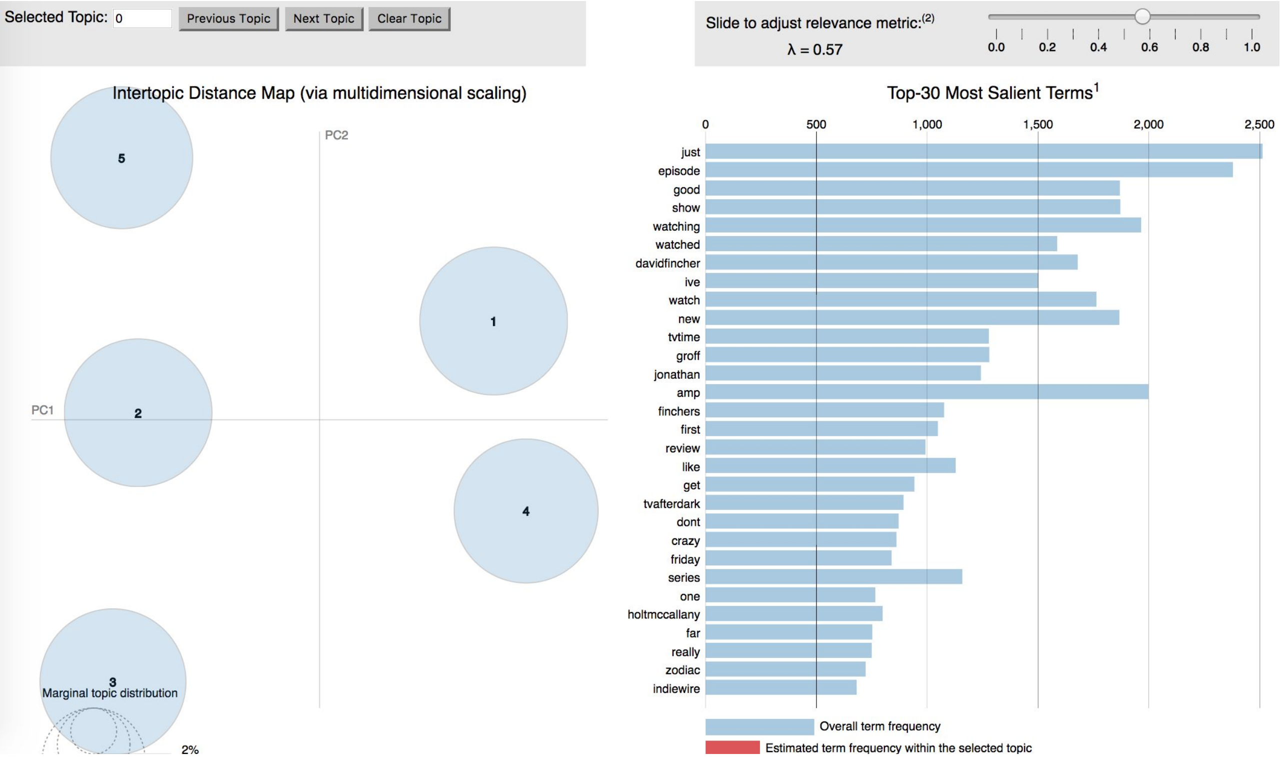
Based on the public Airbnb data: http://insideairbnb.com/get-the-data.html



Things we will create

Topic models

Based on the public Airbnb data: http://insideairbnb.com/get-the-data.html



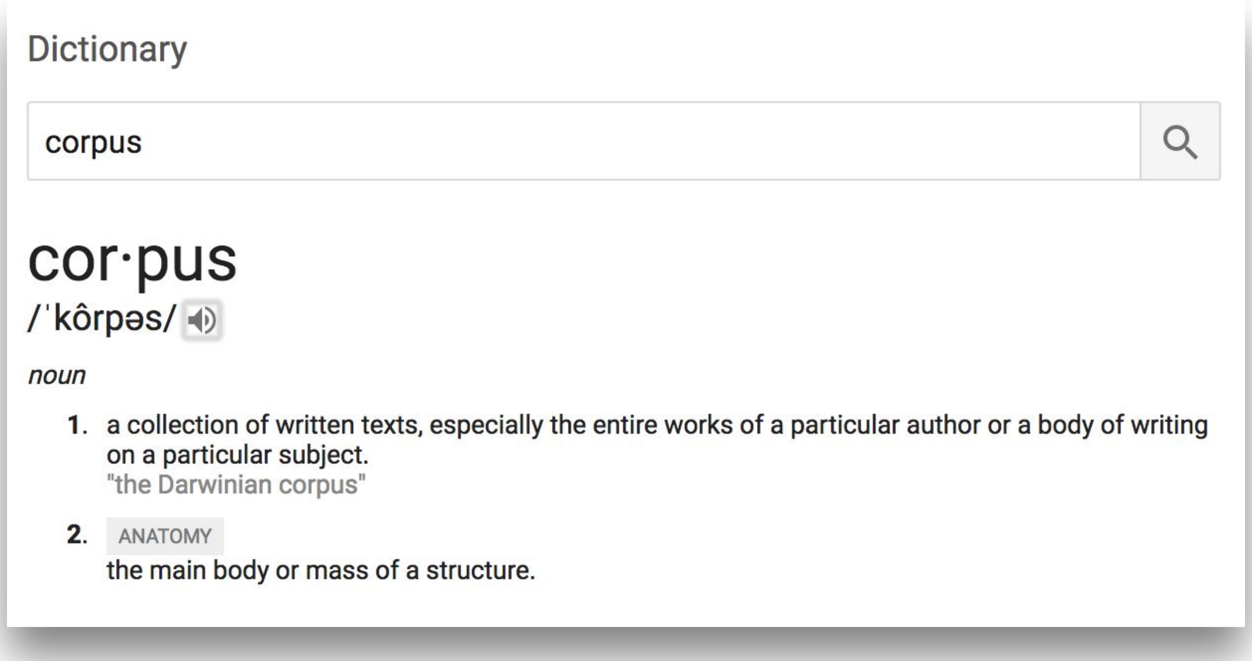
Selected Topic: 0

Slide to adjust relevance metric:(2)

λ = 0.57

Mind the jargons!

- ➔ Corpus
- ➔ Documents



Dictionary

corpus

cor·pus

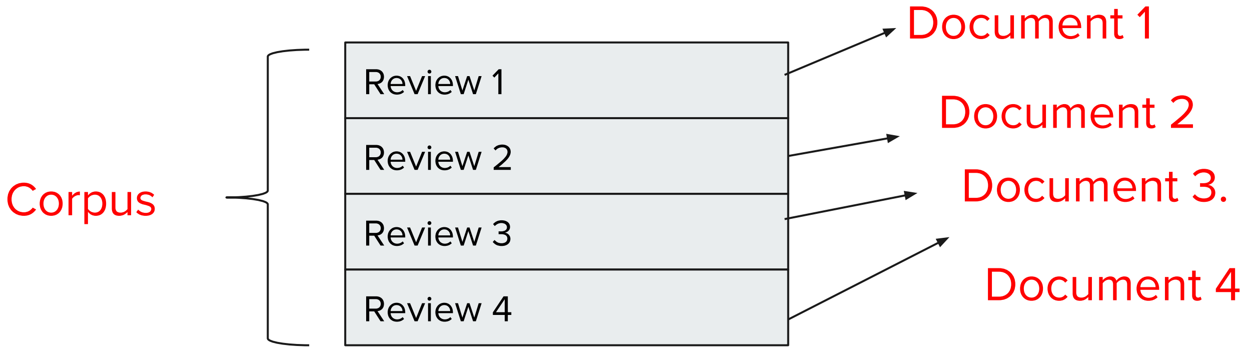
/ˈkôrpəs/

noun

1. a collection of written texts, especially the entire works of a particular author or a body of writing on a particular subject.
"the Darwinian corpus"
2. ANATOMY
the main body or mass of a structure.

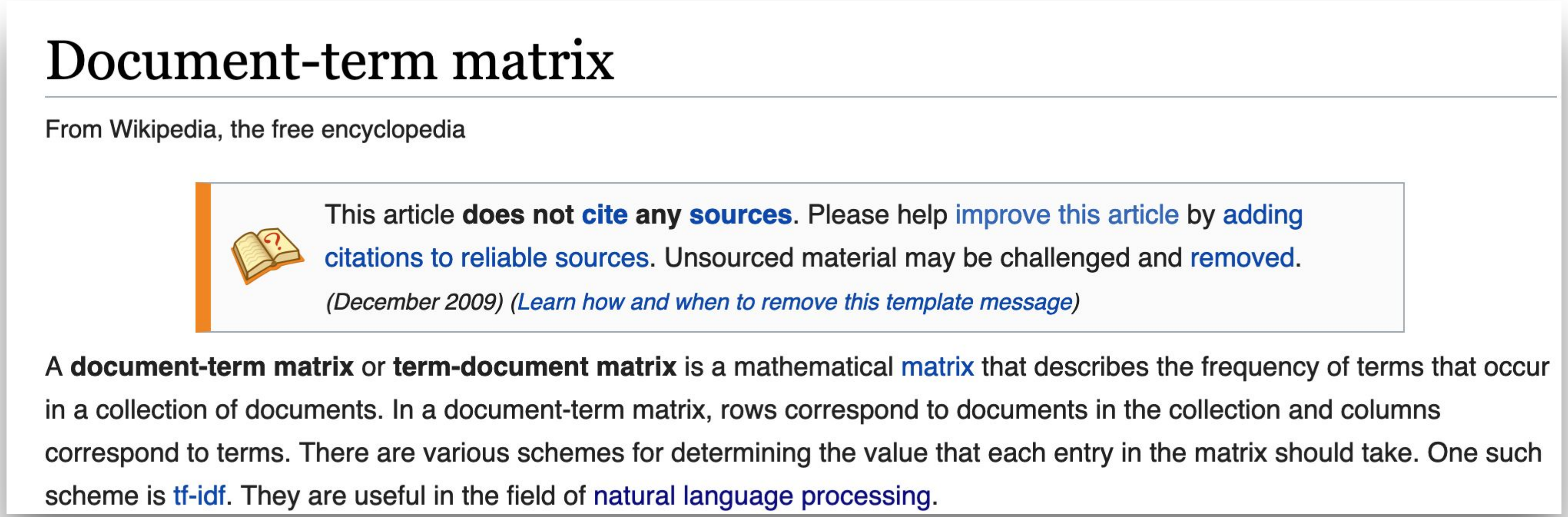
Mind the jargons!

➔ Corpus is a collection of documents



Mind the jargons!

➔ Document-feature matrix (or document-term matrix)



Document-term matrix

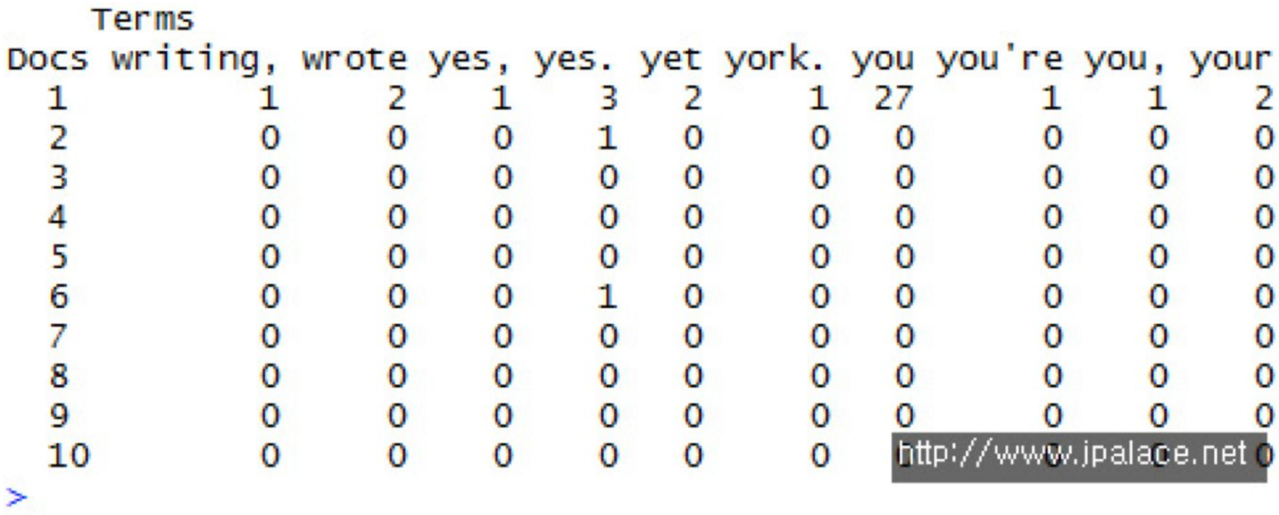
From Wikipedia, the free encyclopedia

This article **does not cite any sources**. Please help improve this article by adding citations to reliable sources. Unsourced material may be challenged and removed.
(December 2009) (Learn how and when to remove this template message)

A **document-term matrix** or **term-document matrix** is a mathematical matrix that describes the frequency of terms that occur in a collection of documents. In a document-term matrix, rows correspond to documents in the collection and columns correspond to terms. There are various schemes for determining the value that each entry in the matrix should take. One such scheme is tf-idf. They are useful in the field of natural language processing.

Mind the jargons!

➔ Document-feature matrix (or document-term matrix)

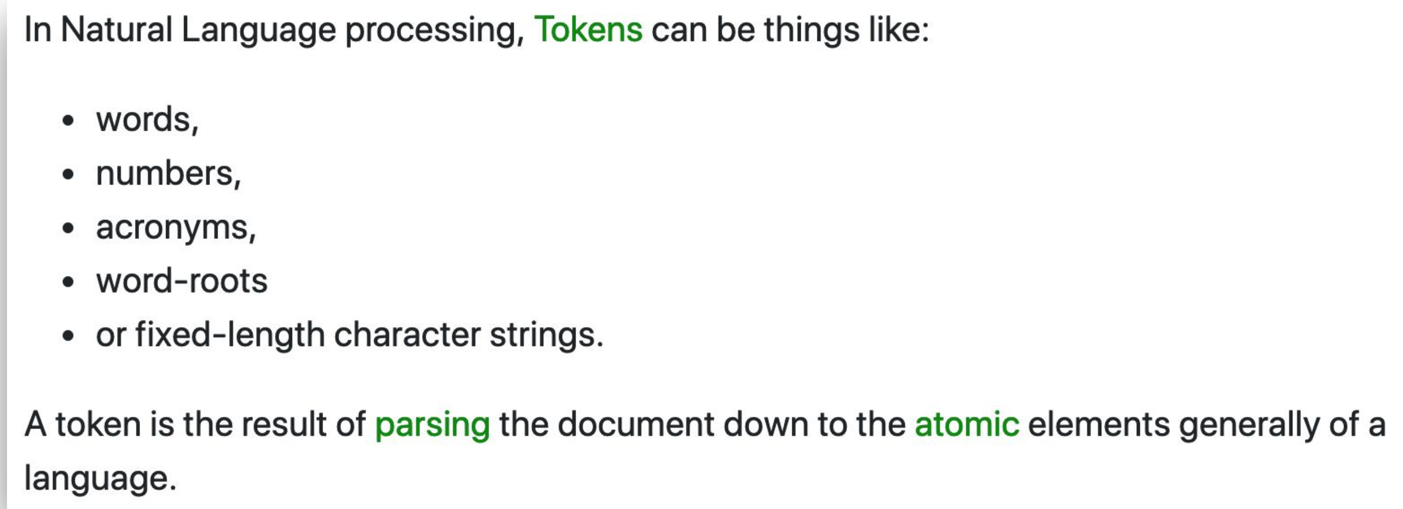


```
    Terms
Docs writing, wrote yes, yes. yet york. you you're you, your
  1         1     2    1    3   2     1  27      1    1    2
  2         0     0    0    1   0     0   0      0    0    0
  3         0     0    0    0   0     0   0      0    0    0
  4         0     0    0    0   0     0   0      0    0    0
  5         0     0    0    0   0     0   0      0    0    0
  6         0     0    0    1   0     0   0      0    0    0
  7         0     0    0    0   0     0   0      0    0    0
  8         0     0    0    0   0     0   0      0    0    0
  9         0     0    0    0   0     0   0      0    0    0
  10        0     0    0    0   0     0   0      0    0    0
>
```

http://www.jpalace.net

Mind the jargons!

➔ Token

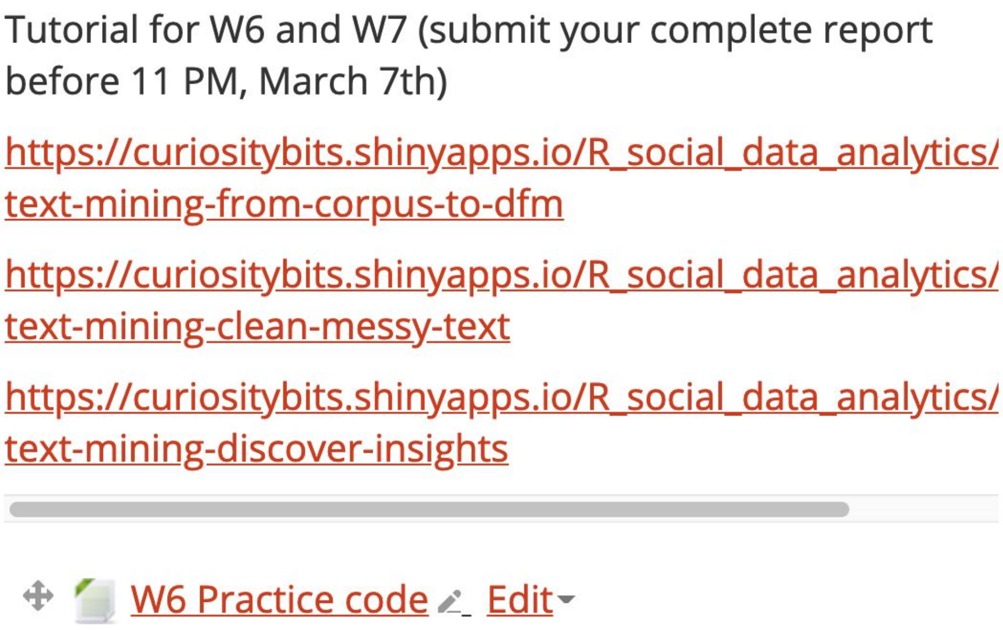


In Natural Language processing, Tokens can be things like:

- words,
- numbers,
- acronyms,
- word-roots
- or fixed-length character strings.

A token is the result of parsing the document down to the atomic elements generally of a language.

Now meet your practice script



Tutorial for W6 and W7 (submit your complete report before 11 PM, March 7th)

https://curiositybits.shinyapps.io/R_social_data_analytics/
text-mining-from-corpus-to-dfm

https://curiositybits.shinyapps.io/R_social_data_analytics/
text-mining-clean-messy-text

https://curiositybits.shinyapps.io/R_social_data_analytics/
text-mining-discover-insights

W6 Practice code