HA1: Collection analysis

WikiIR collection

Source: Wikipedia

We use **en1k** subset of the collection

Size: ~370k documents (cleaned, tokenized, lowercased)
~420 Mb uncompressed

The collection also contains automatically generated queries; we'll use them later.

Description: https://github.com/getalp/wikIR

Example

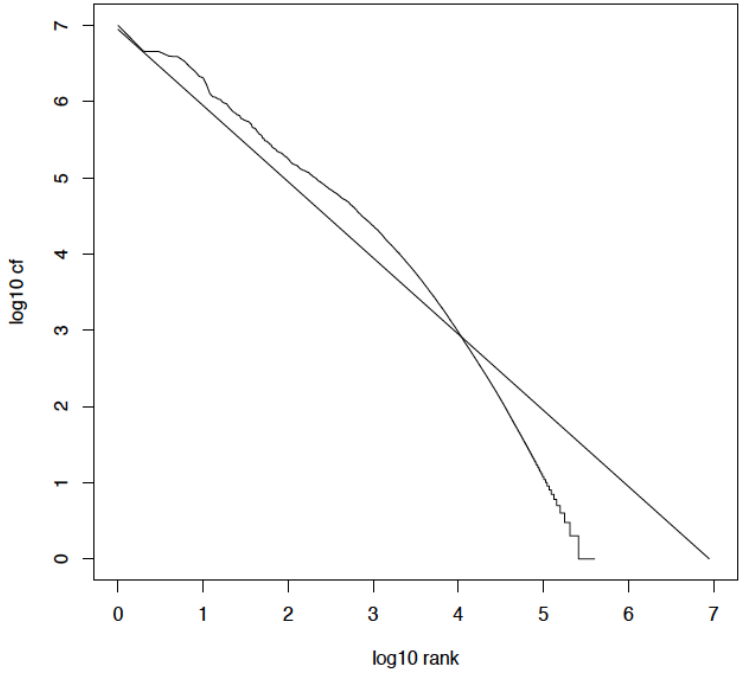
```
1781133,it was used in landing craft during world war ii and is used today in private boats and tra
2426736,after rejecting an offer from cambridge university she moved to london in 1954 working in a
2224122,mat zan coached kuala lumpur fa in 1999 and won malaysia fa cup in 1999 and charity cup in
219642,a barcode is a machine readable optical label that contains information about the item to wh
1728654,since the subordination of the monarchy under parliament and the increasingly democratic me
1889917,the tournament brought together club champions of many domestic beach soccer leagues across
1518473,in january 1948 ballen acquired the interest of goode and became the sole owner he then mov
1459418,it is particularly notable for its association with several of canada s leading blues singe
1102956,they are also known as the wamba wamba wemba wemba bears strong similarities to woiwurrung
931408,they moved south to england to cannock staffordshire where they formed balaam and the angel
1043905,farming and forestry remain central to rural economies and rural development also focuses o
488327,he died while working on the project during the great storm of 1703 he was born in saffron w
1069841,it is a grade i listed building the written history of denton goes back to at least 1253 wh
1601336,she was elected into the house of representatives of nigeria in 2007 to represent ebonyi oh
```

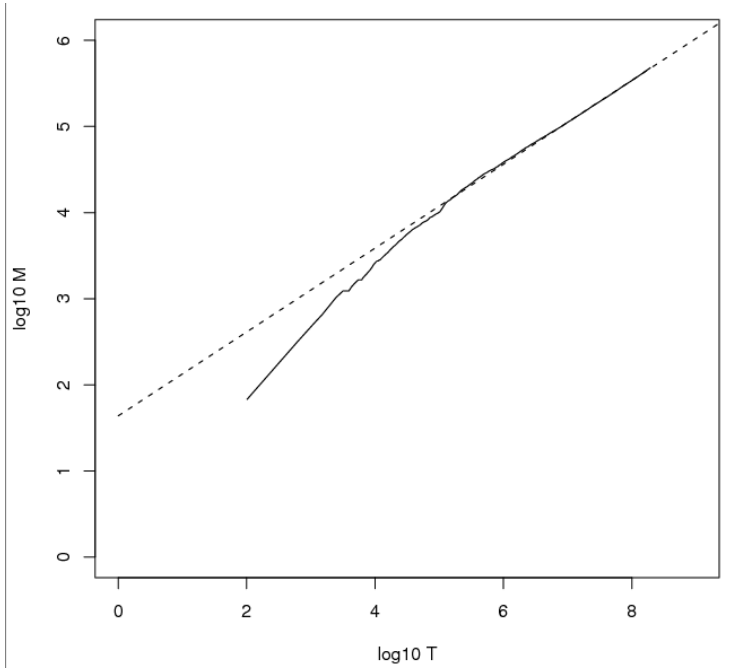
Tasks

- Download collection
- Build a frequency list
  - compare top30 words with a stopword list (e.g. https://gist.github.com/sebleier/554280)
- Provide basic collection stats:
  - # documents
  - Avg. document length in words
  - Collection size in words
  - # unique words
  - Avg. word length
  - Avg. unique word (type) length

Task

- Draw a plot of ranks & collection frequencies in log-log coordinates (Zipf's law)
- Draw a plot of vocabulary growth (unique words) in log-log coordinates (Heaps' law)

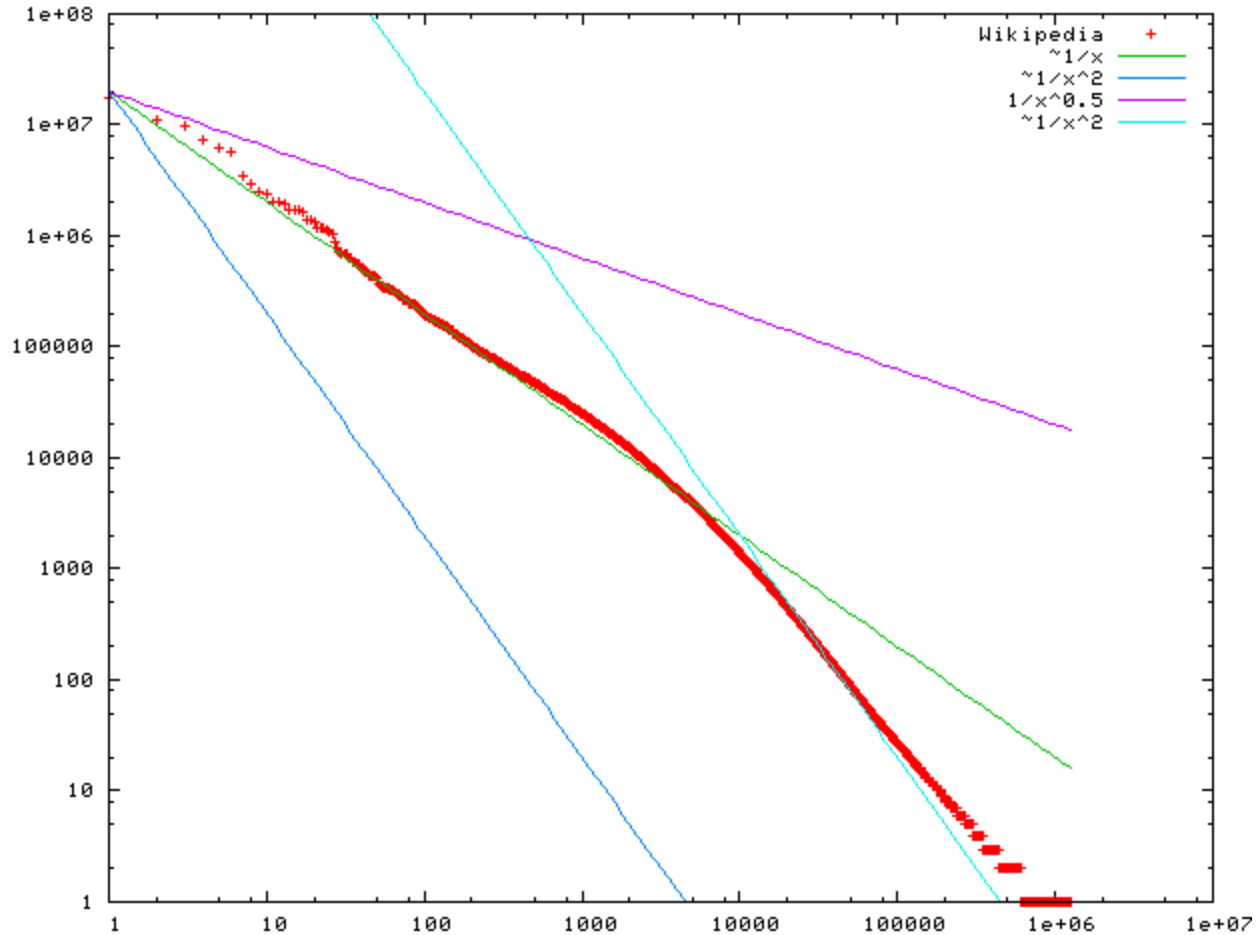




Task *

- Install spaCy https://spacy.io/
- Download trained English model
- *Lemmatize* documents
- Provide basic stats for the lemmatized version of the collection

```
it it
provides provide
information information
technology technology
services service
including include
hosting host
and and
cloud cloud
based base
solutions solution
to to
the the
nhs nhs
and and
many many
other other
organisations organisation
through through
various various
acquisitions acquisition
```


Wikipedia
~1/x
~1/x^2
1/x^0.5
~1/x^2
1e+08
1e+07
1e+06
100000
10000
1000
100
10
1
1
10
100
1000
10000
100000
1e+06
1e+07