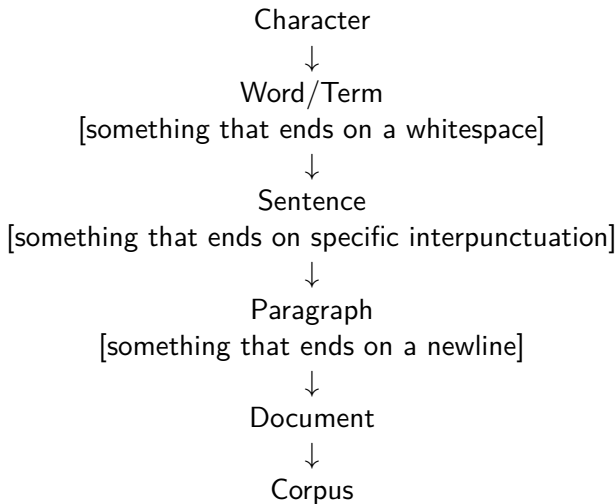


Vectorizing

Examples in Economics

- ▶ Christian Catalini et al. (2015): "[The incidence and role of negative citations in science](#)," Proceedings of the National Academy of Sciences, 112(45).
- ▶ Paul Tetlock (2007): "[Giving Content to Investor Sentiment: The Role of Media in the Stock Market](#)," The Journal of Finance 62(3).
- ▶ Joshua Angrist et al. (2017): "[Economic Research Evolves: Fields and Styles](#)," American Economic Review, 107(5).

Vocabulary on Vocabulary



How to turn documents into numbers

1. Remove stopwords (am, you)
 2. Stem words (drinking → drink)
 3. Tokenize
 4. Remove interpunctuation and numbers
 5. Eventually construct n -grams
 6. Build vocabulary
 7. Encode (words to counts)
- All happens under the hood

What is Vectorization?

- ▶ Turning words into numbers
- ▶ Create $W \times D$ matrix L for W words and D documents
- ▶ $L_{w,d}$ indicates how often document d uses word w
- ▶ Optionally transform the matrix according to tfidf

Vectorizing a document

Document 1:

- ▶ Document 1: burger, ketchup, beer, salad
- ▶ Document 2: kassler, sauerkraut, beer, salad

Vectorizing a document, cont.

Obtain the count matrix:

burger	1	0
ketchup	1	0
beer	1	1
salad	1	1
kassler	0	1
sauerkraut	0	1

Vectorizing a document, cont.

Apply tfidf-transformation:

burger	0.58	0
ketchup	0.58	0
beer	0.41	0.41
salad	0.41	0.41
kassler	0	0.58
sauerkraut	0	0.58

Cosine similarity is $1 - 0.664 \approx 0.336$

tfidf-transformation

tf: term frequency

idf: inverse document frequency

$$\text{tfidf}(w, d) = \underbrace{f_{w,d}}_{\text{tf}} \times \underbrace{\log \left(\frac{D+1}{D_w+1} + 1 \right)}_{\text{idf}}$$

- ▶ $f_{w,d}$: Count of word w in d
- ▶ D : number of documents
- ▶ D_w : number of documents using w

Clustering from Text

- ▶ Vectorization generates matrix
- ▶ You can apply any clustering algorithm on the matrix