Week 9: Data Science Part-Time Course

Natural Language Processing

Dami Lasisi

What is NLP?

- Uses computers to process natural languages
- Tries to make sense of human knowledge stored as unprocessed data
- Builds probabilistic models using data about a language

How is NLP used?

- ‣ Analysis:
  - analyzing positivity and negativity of comments on different websites
  - Extracting key words from text and visualizing how subject topic change over time
- ‣ Vectorizing for Machine Learning
  - Stemming: used for understanding related words
  - Term Frequency-Inverse Document Frequency (TF-IDF): used to identify which world are more likely to be used in the document

NLP High-level task areas

- Chatbots
- Machine translation
- Sentiment Analysis
- Predictive text input
- Natural language generation
- Information retrieval

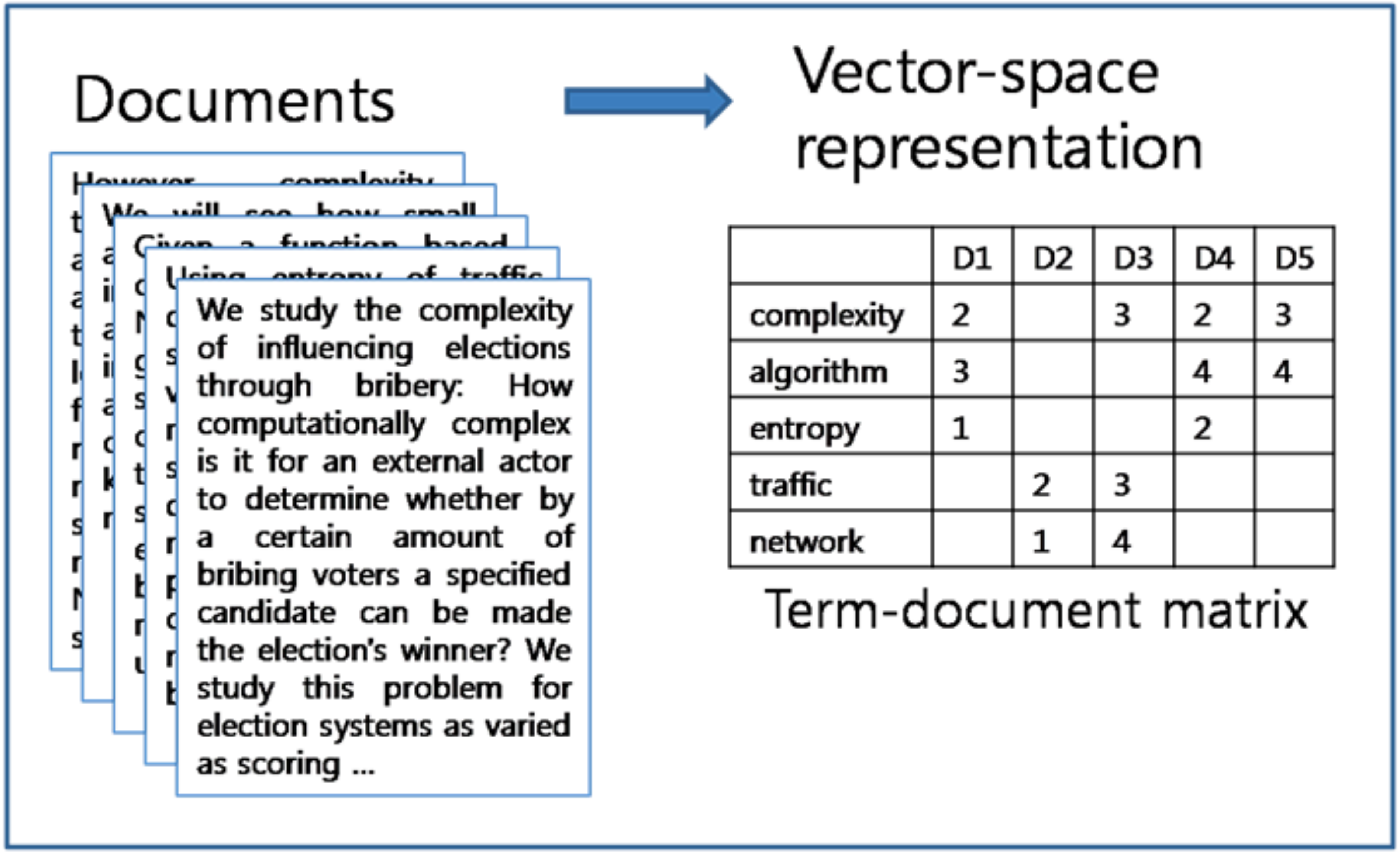
Challenges with NLP

- Ambiguity
- Non-standard text
- Idioms
- Newly coined words
- Tricky entity names

Text Classification

- Text is vectorized into a set of numeric values: each unique word is made a single feature
    - Numeric value of each feature would be the number of times the word appears in the document
    - Make each column an indicator column where 1 would represent the presence of a word and 0 otherwise
- Apply a standard machine learning classifier

Text Classification



Documents

We study the complexity of influencing elections through bribery: How computationally complex is it for an external actor to determine whether by a certain amount of bribing voters a specified candidate can be made the election's winner? We study this problem for election systems as varied as scoring ...

Vector-space representation

	D1	D2	D3	D4	D5
complexity	2		3	2	3
algorithm	3			4	4
entropy	1			2	
traffic		2	3		
network		1	4		

Term-document matrix

Frequently used terms in NLP

- Corpus: a collection of documents
- Corpora: plural form of corpus
- N-grams: features consisting of N consecutive words
- Stop-word removal: process used to remove common words that will likely appear in any text
- Stemming and lemmatization