Computational Communication Science 2
Week 2 - Lecture
Bottom-up Approaches to Text Analysis: From Preprocessing to Vectorization

Anne Kroon
a.c.kroon@uva.nl, @annekroon

April 7, 2025

Digital Society Minor, University of Amsterdam

Today's Agenda

Recap: What We Covered Last Week

Review of Key Concepts

- **Text Preprocessing:** Cleaning and preparing raw text for analysis.
- **Core Techniques:**
  - Stopword removal
  - Stemming and lemmatization
  - Using built-in string methods for text cleaning
- **Methodological Approaches:** Combining *bottom-up* (data-driven) and *top-down* (theory-driven) approaches.

Typical preprocessing steps

Preprocessing steps

tokenization How do we (best) split a sentence into tokens (terms, words)?

pruning How can we remove unneccessary words/ punctuation?

lemmatization and stemming How can we make sure that slight variations of the same word are not counted differently?

ngrams Neighbouring terms

Tokenization and Document-Term Matrix (DTM)

Tokenization: Splitting text into words or subwords.

```python
from nltk.tokenize import TreebankWordTokenizer
tokens = [TreebankWordTokenizer().tokenize(d) for d in docs]
```

Understanding N-grams

- An n-gram is a sequence of *n* words treated as a single feature.
- Examples:
  - Unigrams (1-word units): "science"
  - Bigrams (2-word units): "data science"
  - Trigrams (3-word units): "machine learning model"
- **Why use n-grams?** Captures context and word relationships beyond single words.

Stemming and Lemmatization

- **Stemming:** Reduces words to their base form (e.g., "running" $\rightarrow$ "run").
- **Lemmatization:** Maps words to their dictionary form (e.g., "better" $\rightarrow$ "good").
- Stemming is fast but can produce non-standard words; lemmatization is more accurate but computationally expensive.

```
from nltk.stem import PorterStemmer
stemmer = PorterStemmer()
tokens_stemmed = [stemmer.stem(word) for word in tokens]
```

Exercise time: Word cloud

Let's put this into practice!

- Follow the instructions in the exercise material.
- You will create a word cloud from text data.
- Apply preprocessing techniques such as tokenization, stopword removal, and normalization.

Exercise Link: GitHub: Word cloud exercise

Time: 5 minutes

Tip: Think about how different preprocessing choices impact the

Building text representations: vectorizers

Building text representations: vectorizers

General idea

A text as a collections of word

Let us represent a string

```python
t = "This this is is is a test test test"
# like this:
print(Counter(t.split()))
```

```
Counter({'is': 3, 'test': 3, 'This': 1, 'this': 1, 'a': 1})
```

Compared to the original string, this representation

- is less repetitive
- preserves word frequencies
- but does *not* preserve word order
- can be interpreted as a vector to calculate with (!!!)

Of course, still a lot of stuff to fine-tune... (for example, This/this)

From vector to matrix

If we do this for multiple texts, we can arrange the vectors in a table.

t1 = "This this is is is a test test test"

t2 = "This is an example"

	a	an	example	is	this	This	test
t1	1	0	0	3	1	1	3
t2	0	1	1	1	0	1	0

What can you do with such a matrix? Why would you want to represent a collection of texts in such a way?

What is a vectorizer

- Transforms a list of texts into a sparse (!) matrix (of word frequencies)
- Vectorizer needs to be "fitted" to the training data (learn which words (features) exist in the dataset and assign them to columns in the matrix)
- Vectorizer can then be re-used to transform other datasets

The cell entries: raw counts versus tf·idf scores

- In the example, we entered simple counts (the "term frequency")

But are all terms equally important?

The cell entries: raw counts versus tf·idf scores

- In the example, we entered simple counts (the "term frequency")
- But does a word that occurs in almost all documents contain much information?
- And isn't the presence of a word that occurs in very few documents a pretty strong hint?
- **Solution: Weigh by *the number of documents in which the term occurs at least once) (the "document frequency")***

⇒ we multiply the "term frequency" (tf) by the inverse document frequency (idf)

(usually with some additional logarithmic transformation and normalization applied, see https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfTransformer.html)

TF-IDF: Weighted Importance of Words

- Some words are more informative than others.
- **TF-IDF** adjusts for term frequency and rarity across documents.

$$\text{tf-idf} = \text{tf}_{i,j} \times \log\left(\frac{N}{\text{df}_i}\right)$$

- $\text{tf}_{i,j}$ = term frequency of term i in document j
- df_i = number of documents containing the term i
- N = total number of documents in the corpus

Explanation:

- **Term Frequency (TF)**: Measures how often a term appears

Is tf·idf always better?

It depends.

- Ultimately, it's an empirical question which works better ($\rightarrow$ machine learning)
- In many scenarios, "discounting" too frequent words and "boosting" rare words makes a lot of sense (most frequent words in a text can be highly un-informative)
- Beauty of raw tf counts, though: interpretability + describes document in itself, not in relation to other documents

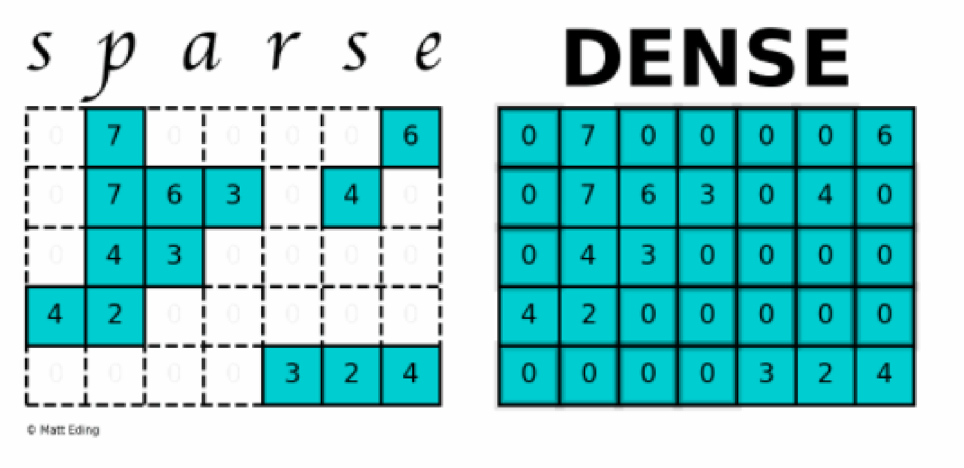
Different vectorizers

1. CountVectorizer (=simple word counts)
2. TfidfVectorizer (word counts ("term frequency") weighted by number of documents in which the word occurs at all ("inverse document frequency"))

Internal representations

Sparse vs dense matrices

- $\rightarrow$ tens of thousands of columns (terms), and one row per document
- Filling all cells is inefficient *and* can make the matrix too large to fit in memory (!!!)
- Solution: store only non-zero values with their coordinates! (sparse matrix)
- dense matrix (or dataframes) not advisable, only for toy examples



https://matteding.github.io/2019/04/25/sparse-matrices/

We learned in week 1 how to tokenize with a list comprehension (and that's often a good idea!).

```
from nltk.tokenize import TreebankWordTokenizer
tokens = [TreebankWordTokenizer().tokenize(d) for d in docs]
```

But what if we want to *directly* get a DTM instead of lists of tokens?

OK, good enough, perfect?

scikit-learn's CountVectorizer (default settings)

- applies lowercasing
- deals with punctuation etc. itself
- minimum word length > 1
- more technically, tokenizes using this regular expression: `r"(?u)\b\w\w+\b"` [1]

```
from sklearn.feature_extraction.text import CountVectorizer
cv = CountVectorizer()
dtm_sparse = cv.fit_transform(docs)
```

[1] ?u = support unicode, \b = word boundary

OK, good enough, perfect?

CountVectorizer supports more

- stopword removal
- custom regular expression
- or even using an external tokenizer
- ngrams instead of unigrams

see
https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html

Best of both worlds

Use the Count vectorizer with a NLTK-based external tokenizer! (see book)

THis notebook might help to better your understanding of vectorizers!

Building text representations: vectorizers

Pruning

General idea

- Idea behind both stopword removal and tf·idf: too frequent words are uninformative
- (possible) downside stopword removal: a priori list, does not take empirical frequencies in dataset into account
- (possible) downside tf·idf: does not reduce number of features

Pruning: remove all features (tokens) that occur in less than X or more than X of the documents

CountVectorizer, only stopword removal

```python
from sklearn.feature_extraction.text import CountVectorizer, TfidfVectorizer
myvectorizer = CountVectorizer(stop_words=mystopwords)
```

CountVectorizer, better tokenization, stopword removal (pay attention that stopword list uses same tokenization!):

```python
myvectorizer = CountVectorizer(tokenizer = TreebankWordTokenizer().tokenize, stop_words=mystopwords)
```

Additionally remove words that occur in more than 75% or less than $n = 2$ documents:

```python
myvectorizer = CountVectorizer(tokenizer = TreebankWordTokenizer().tokenize, stop_words=mystopwords, max_df=.75, min_df=2)
```

All together: tf·idf, explicit stopword removal, pruning

```python
myvectorizer = TfidfVectorizer(tokenizer = TreebankWordTokenizer().tokenize, stop_words=mystopwords, max_df=.75, min_df=2)
```

What is "best"? Which (combination of) techniques to use, and how to decide?

From Text to Features

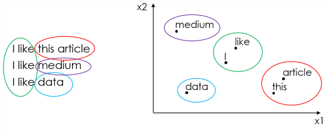
- Text needs to be transformed into numerical representations for computational analysis.
- Tokenization breaks text into meaningful units (words, phrases, n-grams).
- Frequency-based representations allow us to quantify text characteristics.

Introducing embeddings

Why move beyond TF-IDF?

Limitations of TF-IDF and CountVectorizer:

- Create **sparse**, high-dimensional representations.
- Ignore **semantic meaning**—similar words have different vectors.
- Cannot capture **contextual relationships**.



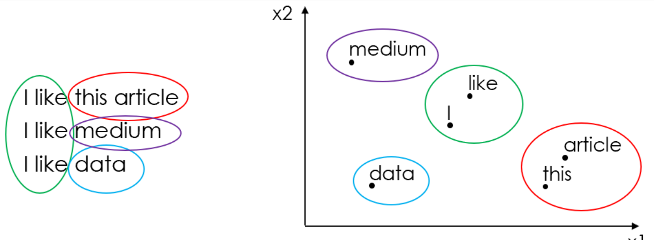
Solution: Word embeddings

Embeddings create **dense**, low-dimensional representations that retain meaning and context.

What are embeddings?

- Map words or documents into **continuous vector spaces**.
- Words with similar meanings have **similar vectors**.
- **Learned** from large text corpora using machine learning models.

Popular Examples: Word2Vec, GloVe, fastText, BERT embeddings.



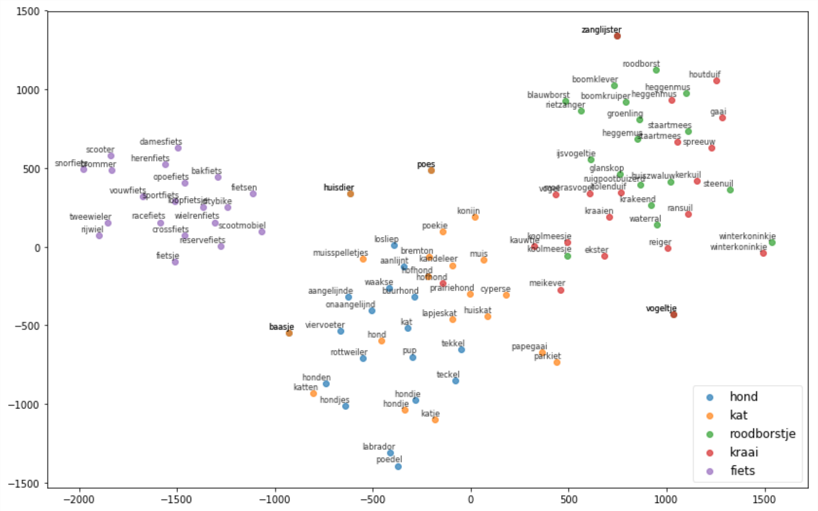
Introducing embeddings

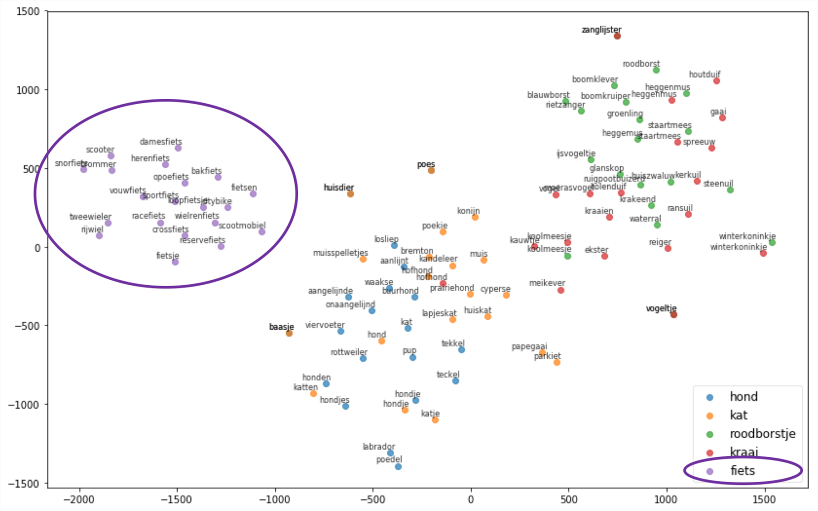
Word embeddings

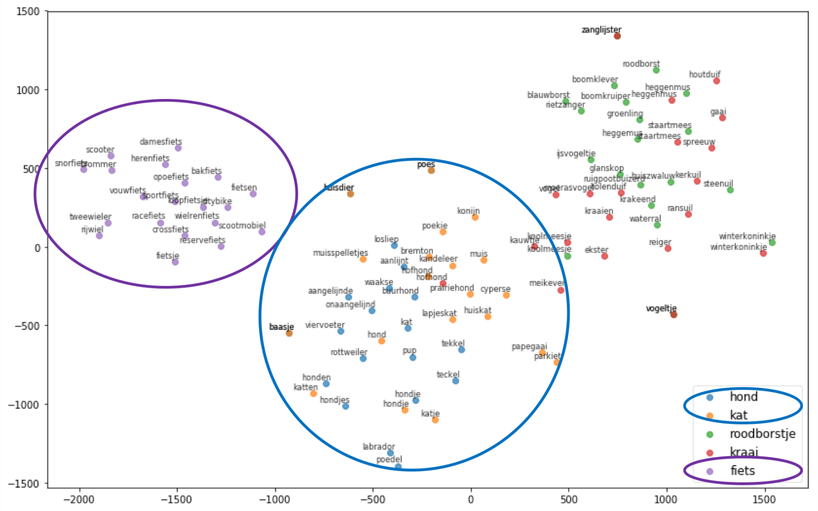
Understanding embeddings

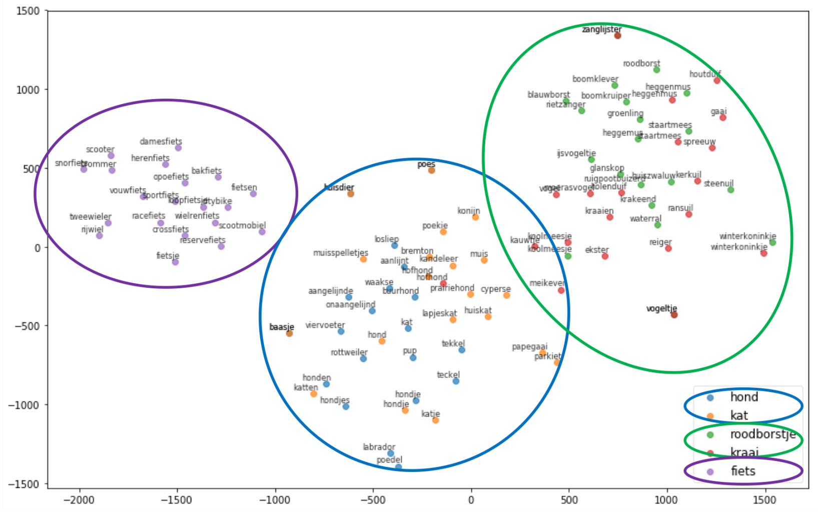
What are word embeddings?

- No technical details here, just the general idea
- Word embeddings help capture the meaning of text
- Word embeddings are low-dimensional vector representations that capture semantic meaning
- Used to be state-of-the-art in NLP (but now: contextualized embeddings, e.g., BERT or GPT)
- *"...a word is characterized by the company it keeps..."* (Firth, 1957)


1500
1000
500
0
−500
−1000
−1500
−2000
−1500
−1000
−500
0
500
1000
1500
zanglijster
roodborst
houtduif
boomklever
heggenmus
blauwborst
boomkruiper
rietzanger
groenling
gaai
staartmees
heggemus
spreeuw
ijsvogeltje
glanskop
kerkuil
steenuil
ruigpootbuizerd
torenduif
krakeend
kraaien
ransuil
waterral
winterkoninkje
reiger
koolmeesje
kauwtje
ekster
meikever
vogeltje
papegaai
parkiet
damesfiets
scooter
herenfiets
snorfiets
brommer
bakfiets
opoefiets
fietsen
vouwfiets
sportfiets
citybike
tweewieler
racefiets
wielrenfiets
rijwiel
crossfiets
scootmobiel
reservefiets
fietsje
poes
huisdier
konijn
poekie
losliep
bremton
muisspelletjes
aanlijnt
kandeleer
muis
hofhond
waakse
prairiehond
cyperse
aangelijnde
buurhond
onaangelijnd
lapjeskat
huiskat
baasje
viervoeter
kat
hond
tekkel
rottweiler
pup
teckel
honden
katten
hondjes
hondje
katje
labrador
poedel
hond
kat
roodborstje
kraai
fiets


1500
1000
500
0
−500
−1000
−1500
−2000
−1500
−1000
−500
0
500
1000
1500
zanglijster
roodborst
houtduif
boomklever
heggenmus
blauwborst
boomkruiper
rietzanger
groenling
gaai
staartmees
heggemus
spreeuw
ijsvogeltje
damesfiets
scooter
herenfiets
snorfiets
brommer
poes
glanskop
bakfiets
opoefiets
kerkuil
steenuil
fietsen
huisdier
vouwfiets
sportfiets
loopfiets
citybike
krakeend
kraaien
ransuil
tweewieler
racefiets
wielrenfiets
konijn
waterral
rijwiel
crossfiets
scootmobiel
poekie
reservefiets
losliep
koolmeesje
winterkoninkje
fietsje
bremton
reiger
muisspelletjes
aanlijnt
kandeleer
muis
ekster
hofhond
waakse
prairiehond
cyperse
meikever
aangelijnde
buurhond
onaangelijnd
lapjeskat
huiskat
vogeltje
kat
baasje
viervoeter
hond
tekkel
papegaai
rottweiler
pup
parkiet
teckel
honden
katten
hondje
hondjes
katje
labrador
poedel
hond
kat
roodborstje
kraai
fiets


hond
kat
roodborstje
kraai
fiets
zanglijster
poes
huisdier
baasje
vogeltje
damesfiets
herenfiets
bakfiets
fietsen
scooter
brommer
snorfiets
vouwfiets
sportfiets
racefiets
tweewieler
rijwiel
crossfiets
reservefiets
fietsje
wielrenfiets
scootmobiel
roodborst
houtduif
boomklever
heggenmus
blauwborst
rietzanger
groenling
gaai
staartmees
spreeuw
ijsvogeltje
glanskop
kerkuil
steenuil
krakeend
kraaien
waterral
ransuil
reiger
ekster
winterkoninkje
meikever
konijn
poekie
losliep
bremton
muis
muisspelletjes
aanlijnt
kandeleer
hofhond
waakse
prairiehond
cyperse
aangelijnde
onaangelijnd
lapjeskat
huiskat
viervoeter
kat
hond
tekkel
papegaai
parkiet
rottweiler
pup
teckel
honden
katten
hondjes
hondje
katje
labrador
poedel



- Explore word embeddings interactively using the TensorFlow Projector.
- Click here to access the TensorFlow Projector.

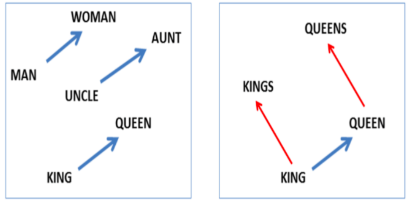
How embeddings work

- Each word (or document) is represented by a **vector** in a high-dimensional space.
- The model learns these vectors by predicting word co-occurrences in text.
- Example: **Word2Vec** uses a neural network to predict surrounding words.

Example: Word similarity with embeddings

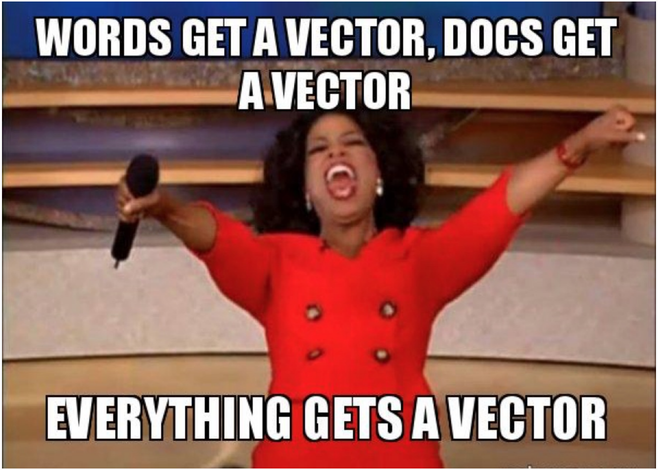
Word2Vec analogy:

- `vec("king") - vec("man") + vec("woman") = vec("queen")`
- Captures **semantic relationships** automatically.
- Unlike TF-IDF, embeddings **understand meaning**.



Example: Getting word embeddings with spaCy

Try it out yourself.. Visualize your embeddings



Embeddings in communication science

Applications:

- **Sentiment analysis:** Understanding audience reactions on social media (Rudkowsky et al., 2018)
- **Topic modeling**: cross lingual topics (**han2020reproducible**)
- **Recommender Systems:** Suggesting content based on user preferences (Loecherbach and Trilling, 2020)
- **Text similarity:** Measuring the similarity between texts.
  - Example: Brinberg and Ram, 2021

Note: We will discuss this next week.

Using pretrained embeddings

Why use pretrained models?

- Trained on massive datasets (Google News, Wikipedia, etc.).
- Capture **rich linguistic structures**.
- Reduce training time and improve performance.

Popular Choices:

- Word2Vec (Google News)
- GloVe (Common Crawl, Wikipedia)
- BERT (contextualized embeddings)
- fastText (subword information)

Why Use Embeddings?

- Capture **semantic meaning** of words.
- Handle **synonyms** and **related words** effectively.
- Work well in NLP applications: **text classification, clustering, sentiment analysis**.
- Used in communication science for **media analysis, misinformation detection, and social network studies**.

Key takeaways

- Traditional methods (TF-IDF) are **limited in capturing meaning**.
- Word embeddings create **dense vectors** that capture relationships.
- Pretrained models like **Word2Vec, GloVe, and BERT** help analyze text effectively.
- Embeddings are **widely used** in NLP and communication science.