Introduction to Text Mining and Natural Language Processing

Session 3: Text Mining Pipeline and the DTM

Hannes Mueller

January 2026

Barcelona School of Economics

Orientation

Course Overview

Part 1: Project Design and Getting the Text (Session 2)

Part 2: Text Mining Basics (Sessions 3 and 4)

Part 3: Dimensionality Reduction (Sessions 5 to 7)

In-class assignment in session 6

Part 4: Supervised Learning with Text (Sessions 8 and 9)

In-class assignment in session 9

Write me your term-paper ideas throughout.

Term paper presentations: 16th of March (feedback!)

This session

Discussion of last TA - DiD

Preprocessing - getting to the document term matrix

Discuss with your neighbor

VARIABLES	(1) price	(2) lnprice	(3) lnprice	(4) lnprice
treatment	705.768***	0.227***	0.227***	0.175**
	(86.133)	(0.017)	(0.053)	(0.069)
Observations	20,755	20,755	20,755	19,384
R-squared	0.077	0.360	0.360	0.952
FE	City + Date	City + Date	City + Date	Hotel + Date
SE	Robust (HC1)	Robust (HC1)	Cluster(city)	Cluster(city)

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Pre-Processing Overview

Introduction to Pre-processing

- Pre-processing is often done "before" we start to apply another method.
- But, in a way this is not really a separate step to the analysis.
- Very often the analysis is really determined in the collection and pre-processing stages.
- Before you get the data you need to have some idea of why you are mining.

Reminder Definitions

- A single observation in a textual database is called a *document*, $d \in \{1, ..., D\}$.
- The set of documents that make up the dataset is called a *corpus* D.
- We often have covariates associated with each document that are sometimes called *metadata*.
- When processing the first step is called *tokenization*. Tokens include punctuation and can, depending on the method, even be parts of words.
- The vocabulary V is spanned by terms v.

Our Workhorse: Bag-of-Words Model

- Bag-of-words model means we lose grammar - order does not matter.
- A *document*, d can then be represented by a count of *terms*, v.
- Example: *The president of the United States is Joe Biden.*
- Document vector, d after transformation in unigrams:

the	*president*	*of*	*united*	*states*	*is*	*joe*	*biden*
2	1	1	1	1	1	1	1

- Note difference between terms and *words* in this model.

Result of Pre-Processing

- After pre-processing, each document is a finite list of terms.
- A basic representation of a corpus is the following:
  - Index each unique term in the corpus by some $v \in \{1, \ldots, V\}$, where V is the number of unique terms.
  - For each document $d \in \{1, \ldots, D\}$ compute the count $x_{d,v}$ as the number of occurrences of term v in document d.
- The $D \times V$ matrix X of all such counts is called the document-term matrix.
- This representation is often called the bag-of-words model.

Example of document-term matrix

Having an example matrix will be useful. Imagine you have three documents (in this case sentences)

	$v=1$	$v=2$	$v=3$	$v=4$	$v=5$	$v=6$	$v=7$	$v=8$
	the	president	of	united	states	is	joe	biden
d=1	2	1	1	1	1	1	1	1
d=2	0	1	0	0	0	1	1	1
d=3	1	0	0	1	1	0	0	0

Full DTM notation

$$X = \begin{bmatrix} x_{11} & x_{12} & x_{13} & x_{14} & x_{15} & \cdots & x_{1V} \\ x_{21} & x_{22} & x_{23} & x_{24} & x_{25} & \cdots & x_{2V} \\ x_{31} & x_{32} & x_{33} & x_{34} & x_{35} & \cdots & x_{3V} \\ x_{41} & x_{42} & x_{43} & x_{44} & x_{45} & \cdots & x_{4V} \\ x_{51} & x_{52} & x_{53} & x_{54} & x_{55} & \cdots & x_{5V} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{D1} & x_{D2} & x_{D3} & x_{D4} & x_{D5} & \cdots & x_{DV} \end{bmatrix}$$

This DTM is a $D \times V$ matrix.

N-grams

- The vocabulary of the DTM is set in a step called vectorization.
- In that step we often want to build terms v from word combinations
- Combinations of more than one word in a row are called N-grams. For example, bigrams, trigrams...
- Common n-grams often carry offices concepts or organizations
- Typical bigrams in the example above will be *united_states* and *joe_biden*

Pre-processing Overview

Steps to create a DTM:

- Tokenization: Splitting text into tokens (last session).
- Normalization: Doing things to the text to reduce dimensionality (this session)
- Vectorization - we build the DTM (this session):
  - Dictionary: Build the dictionary of all terms.
  - Counting: Counting the frequency of each term in each document.
  - Organizing: Structuring the data into the matrix.

Normalization

Steps to follow (typically):

- lemmatizing
- remove punctuation
- unify, e.g. convert to lower case
- remove stopwords
- stemming

Stemming/Lemmatizing

- Stemming: Deterministic algorithm for removing suffixes. Porter stemmer is popular.
  - In the Porter algorithm, argue, argued, argues, arguing, and argus reduce to the stem argu. Sometimes equivalence between tokens is misleading: 'university' and 'universe' stemmed to same form.
- Lemmatizing: Tag each token with its part of speech, then look up each (word, POS) pair in a dictionary to find linguistic root. E.g. 'saw' tagged as verb would be converted to 'see', 'saw' tagged as noun left unchanged.
- The following is an example of the middle part of the process.

Text on Israel from 2016

Palestinian President Mahmoud Abbas said Wednesday that late former Israeli President Shimon Peres had exerted unremitting efforts to make peace until the last moment of his life. "He (Peres) exerted unremitting efforts to reach a permanent peace since Oslo agreement was signed with Israel in 1993 until the last moment in his life," Abbas said in a condolence letter to Peres' family, carried by state news agency WAFA. In his letter, Abbas expressed deep grief and sorrow for Peres' passing. "He was a partner in making the peace of the brave with late Palestinian leader Yasser Arafat and late Israeli Prime Minister Yitzhak Rabin." Earlier on Wednesday, Israel announced that Peres died in a hospital in the suburbs of Tel Aviv at the age of 93. "We are so happy about the news of former Israeli president's death. It is not only us, but all the Palestinian people are happy about the news of Peres' death," he said. "This man committed crimes and shed the blood of our people." Abu Zuhri, whose movement is classified as a terrorist group and rejects to recognize Israel and the peace process, said: "Peres was the last founder of this entity (Israel) and we believe it is a start of a new stage of the

Stopwords highlighted

Palestinian President Mahmoud Abbas said Wednesday that late former Israeli President Shimon Peres had exerted unremitting efforts to make peace until the last moment of his life. "He (Peres) exerted unremitting efforts to reach a permanent peace since Oslo agreement was signed with Israel in 1993 until the last moment in his life," Abbas said in a condolence letter to Peres' family, carried by state news agency WAFA. In his letter, Abbas expressed deep grief and sorrow for Peres' passing. "He was a partner in making the peace of the brave with late Palestinian leader Yasser Arafat and late Israeli Prime Minister Yitzhak Rabin." Earlier on Wednesday, Israel announced that Peres died in a hospital in the suburbs of Tel Aviv at the age of 93. "We are so happy about the news of former Israeli president's death. It is not only us, but all the Palestinian people are happy about the news of Peres' death," he said. "This man committed crimes and shed the blood of our people." Abu Zuhri, whose movement is classified as a terrorist group and rejects to recognize Israel and the peace process, said: "Peres was the last founder of this entity (Israel) and we believe it is a start of a new stage of the

Lowercased, no punctuation, stopwords removed

palestinian president mahmoud abbas said wednesday late former israeli president shimon peres exerted unremitting efforts make peace last moment life peres exerted unremitting efforts reach permanent peace since oslo agreement signed israel 1993 last moment life abbas said condolence letter peres family carried state news agency wafa letter abbas expressed deep grief sorrow peres passing partner making peace brave late palestinian leader yasser arafat late israeli prime minister yitzhak rabin earlier wednesday israel announced peres died hospital suburbs tel aviv age 93 happy news former israeli presidents death us palestinian people happy news peres death said man committed crimes shed blood people abu zuhri whose movement classified terrorist group rejects recognize israel peace process said peres last founder entity israel believe start new stage israeli occupations weakness

Stemmed text, no punctuation, stopwords removed

palestinian president mahmoud abba said wednesday late former isra president shimon pere exert unremit effort make peac last moment life pere exert unremit effort reach perman peac sinc oslo agreement sign israel 1993 last moment life abba said condol letter pere famili carri state news agenc wafa letter abba express deep grief sorrow pere pass partner make peac brave late palestinian leader yasser arafat late isra prime minist yitzhak rabin earlier wednesday israel announc pere die hospit suburb tel aviv age 93 happi news former isra presid death us palestinian peopl happi news pere death said man commit crime shed blood peopl abu zuhri whose movement classifi terrorist group reject recogn israel peac process said pere last founder entiti israel believ start new stage isra occup weak

Lemmatized text, no punctuation, stopwords removed

palestinian President Mahmoud Abbas say Wednesday late israeli President Shimon Peres exert unremitting effort peace moment life Peres exert unremitting effort reach permanent peace Oslo agreement sign Israel 1993 moment life Abbas say condolence letter Peres family carry state news agency WAFA letter Abbas express deep grief sorrow Peres pass partner make peace brave late palestinian leader yasser arafat late israeli Prime Minister Yitzhak Rabin early Wednesday Israel announce Peres die hospital suburb Tel Aviv age 93 happy news israeli President death Palestinian people happy news Peres death say man commit crime shed blood people Abu Zuhri movement classify terrorist group reject recognize Israel peace process say Peres founder entity Israel believe start new stage israeli occupation weakness

Vectorizing: Building the Dictionary

After pre-processing, each document is a finite list of terms (unigrams, bigrams...). This gives us a vocabulary. In the example above this looks something like this:

$\{'palestinian' : 0, 'president' : 1, 'mahmoud' : 2, 'abbas' : 3, 'said' : 4, 'wednesday' : 5, 'late' : 6, 'former' : 7, 'israeli' : 8, 'shimon' : 9, 'peres' : 10, 'exerted' : 11, 'unremitting' : 12, 'efforts' : 13, 'make' : 14, 'peace' : 15, 'last' : 16, 'moment' : 17, 'life' : 18, 'reach' : 19, 'permanent' : 20, 'since' : 21, 'oslo' : 22, 'agreement' : 23, 'signed' : 24, 'israel' : 25, '1993' : 26, 'condolence' : 27, 'letter' : 28, 'family' : 29, 'carried' : 30, 'state' : 31, 'news' : 32, 'agency' : 33, ...$

Vectorization: Building the DTM

We then use this vocabulary to get to the corpus representation as the document term matrix:

- Index each unique term in the corpus by some $v \in \{1, ..., V\}$, where V is the number of unique terms.
- For each document $d \in \{1, ..., D\}$ compute the count $x_{d,v}$ as the number of occurrences of term v in document d.

Introduction to CountVectorizer

Command Syntax

```
CountVectorizer(ngram_range=(1, 3),
                min_df=0.05, max_df=0.3)
```

Parameters:

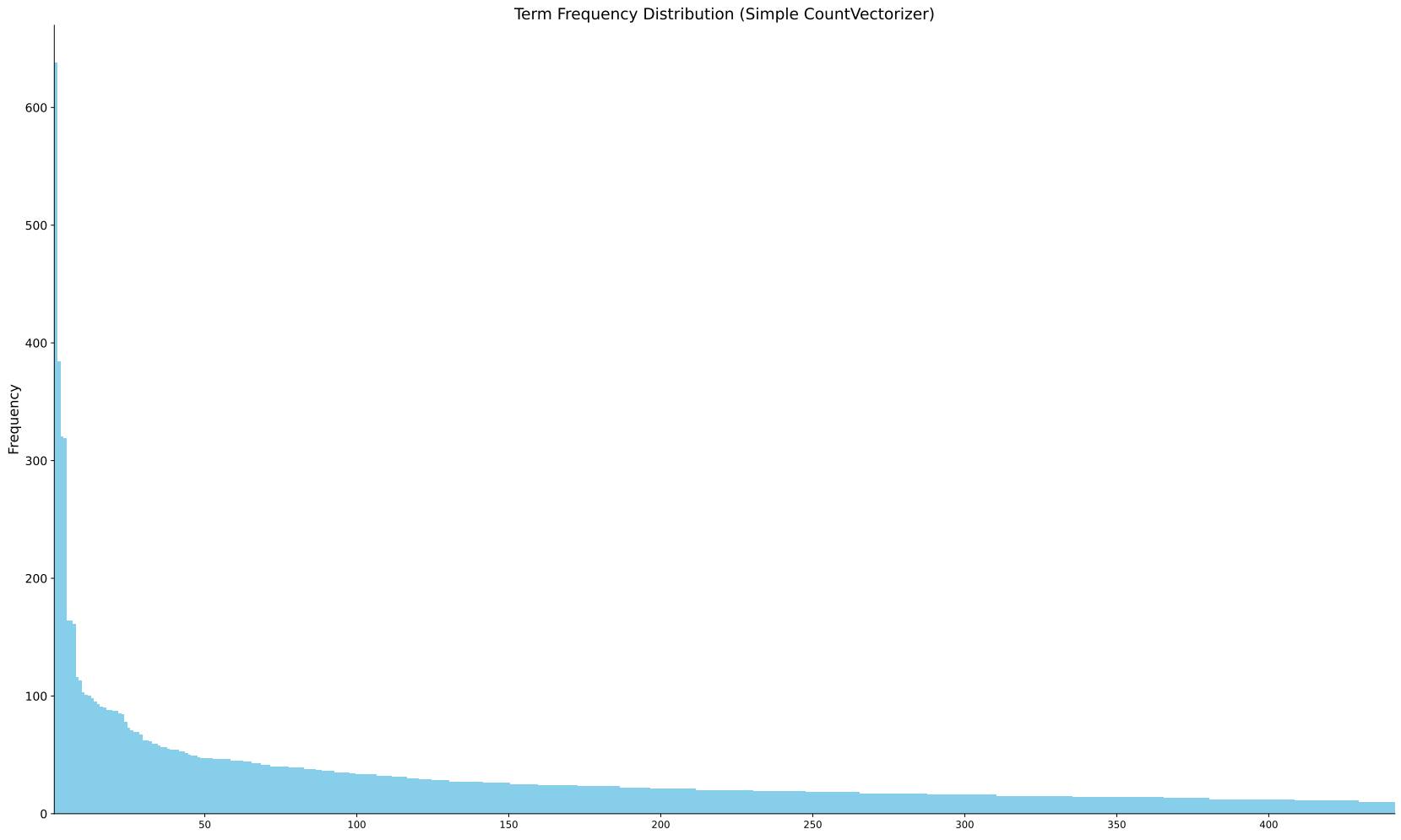
- **ngram_range=(1, 3):** Considers unigrams, bigrams, and trigrams.
- **min_df=0.05:** Ignores terms appearing in fewer than 5% of documents.
- **max_df=0.3:** Ignores terms appearing in more than 30% of documents.

Think about what this does to the term distribution.

Term Frequency Distribution

Let's use min_df to cut off tail.

Term Frequency Distribution (min_df used)



Inputs and Outputs of CountVectorizer

Input

List of text documents (strings).

Outputs

Scipy sparse matrix: each row represents a document, and each column represents a term from the vocabulary. The values indicate the frequency of the term in the document.

Vocabulary: maps each term to its corresponding feature index in the document-term matrix.

Method: `get_feature_names_out()` retrieves the list of terms in the vocabulary, ordered by their feature indices.

Text as Vectors: the Fun Begins

The Matrix View on Text as Data

- We can get some representation of a document through term frequencies, $x_{d,v}$. Three common problems with this:
- 1) some documents are much longer than other documents. What should we do?
- Normalize counts by document length.
  - Let $x_{d,v}$ be the count of the term v in document d.
  - Calculate $f_{d,v} = x_{d,v}/(\sum_v x_{d,v})$
  - Keep this in mind also when aggregating documents!
- 2) some terms that are very common across the corpus.
- 3) sometimes we want to weigh specific content stronger

The Project Design View

Two callenges:

- The number of terms, V, is relatively high. WE need to **reduce dimensionality**:
  - Often you have strong priors on what to use, i.e. specific keywords.
  - Sometimes you know what "kind" of text you want capture, i.e. text which talks about finance.
  - In other applications there are no priors.
- Once you have the features X you typically **need to aggregate** to the level of analysis you are at with your y or other features X.