Introduction to Text Analysis

Arieda Muço

Central European University

March 17, 2022

"Usual" Data

- A spreadsheet with continuous and discrete variables (ready for analysis!?), fixed number of columns
  - Real data is messy and almost never ready for analysis
- Data can come from images, audio, video, text and there is no (fixed) number of columns

Common data

```
<DOC>
<DOCNO>106-biden-de-5-20001215</DOCNO>
<TEXT>
   Mr. BIDEN. We should do that. I get the feeling--maybe because it is the Christmas season and I want to believe it--there is a growing recognition that rail
service in our neck of the woods, as well as other parts of the country, are as essential to our interests as water is to the far west. It is as essential.
   I thank my colleagues for their commitment and absolutely close by saying to Senator BYRD that I appreciate the fact that he understands, maybe better than
anyone in this place, when another colleague cares about an issue that he believes is absolutely indispensable for his region. I thank him for acknowledging that.
   I thank him for his--it is no new commitment; he has always been committed to Amtrak--acknowledgment of that and for his continued pledge of commitment to
Amtrak. With this combination of the majority leader, the Democratic leader, the chairman of the Appropriations Committee, the ranking member of the Appropriations
Committee, and the ranking member of the Commerce Committee, if we cannot get it done, then shame on us.
   I thank all of my colleagues. Sorry to have taken so much time, but as my colleagues said all day, this is a big, big, big, big deal to me personally, to my
State, and I think to the Nation.
   I yield the floor.
</TEXT>
</DOC>
```

13. A LAKÓHELY VÁLTOZTATÁSOK ADATAI

1985

Város, városi jogu nagyközség	Állandó vándorlások		Ideiglenes vándorlások és visszavándorlások		A népesség növekedése /+/, illetve csökkenése /-/		
	odavándorlások	elvándorlások	odavándorlások	elvándorlások	az állandó	az ideiglenes és vissza-	a belföldi
					vándorlások következtében		
MEGYESZÉKHELYEK							
Budapest	25212	14746	83254	78883	+ 10466	+ 4371	+14837
Békéscsaba	1242	1084	2568	2437	+ 158	+ 131	+ 289
Debrecen	4346	2983	10858	9806	+ 1363	+ 1052	+ 2415
Eger	1302	1096	3400	3211	+ 206	+ 189	+ 395
Győr	2125	1696	5080	4866	+ 429	+ 214	+ 643
Kaposvár	1649	1516	2935	2777	+ 133	+ 158	+ 291
Kecskemét	2046	1442	4245	3978	+ 604	+ 267	+ 871
Miskolc	3422	3653	10035	9689	– 231	+ 346	+ 115
Nyiregyháza	2466	1903	4988	5669	+ 563	– 681	– 118
Pécs	3384	2766	9489	8343	+ 618	+ 1146	+ 1764
Salgótarján	952	752	1546	1821	+ 200	– 275	– 75
Szeged	3466	2042	10706	8587	+ 1424	+ 2119	+ 3543
Szekszárd	1010	880	1809	1813	+ 130	– 4	+ 126
Székesfehérvár	2369	1874	5034	4557	+ 495	+ 477	+ 972
Szolnok	1977	1733	4401	3977	+ 244	+ 424	+ 668
Szombathely	1458	1312	3184	3245	+ 146	– 61	+ 85
Tatabánya	1305	1584	2985	3219	– 279	– 234	– 513
Veszprém	1781	1396	3078	2754	+ 385	+ 324	+ 709
Zalaegerszeg	1311	746	2174	2254	+ 565	– 80	+ 485
TÖBBI VAROS							

Also common data

id	nr_programs	nr_pages	tot_fiscaliz	nr_fiscaliz	abb_uf	municipio	nsorteio
	.	.	.	.	BA	lauro de freitas	102
	.	.	.	.	PI	pedro ii	101
01-AM-Rio_Preto_da_Eva	.	11	.	.	AM	rio preto da eva	1
01-GO-Castelandia	.	13	.	.	GO	castelandia	1
01-PI-Colonia_do_Piaui	.	12	.	.	PI	colonia do piaui	1
01-SC-Balneario_Arroio_do_Silva	.	8	.	.	SC	balneario arroio do silva	1
01-SP-Ribeirao_Corrente	.	10	.	.	SP	ribeirao corrente	1
02-AC-Marechal_Thaumaturgo	76	78	2555313	.	AC	marechal thaumaturgo	2
02-AL-Japaratinga	74	61	2234636	74	AL	japaratinga	2
02-AM-Alvaraes	68	64	2371811	68	AM	alvaraes	2
02-AP-Pracuuba	44	33	543222.3	44	AP	pracuuba	2
02-BA-Presidente_Tancredo_Neves	106	52	7978503	106	BA	presidente tancredo neves	2
02-CE-Santa_Quiteria	89	94	6232543	89	CE	santa quiteria	2
02-ES-Jaguare	118	48	3788544	118	ES	jaguare	2
02-GO-Inaciolandia	68	28	.	.	GO	inaciolandia	2
02-MA-Apicum_Acu	93	81	2649186	121	MA	apicum acu	2
02-MG-Sao_Joao_das_Missoes	86	85	33292.78	.	MG	sao joao das missoes	2
02-MS-Vicentina	81	88	2337215	81	MS	vicentina	2
02-MT-Pontal_do_Araguaia	73	75	2083512	73	MT	pontal do araguaia	2
02-PA-Abel_Figueiredo	76	49	1688877	76	PA	abel figueiredo	2
02-PB-Pitimbu	116	116	4833256	116	PB	pitimbu	2
02-PE-Alagoinha	110	42	4833256	116	PE	alagoinha	2
02-PI-Alvorada_do_Gurgueia	87	46	2289395	78	PI	alvorada do gurgueia	2
02-PR-Prudentopolis	129	75	7448459	129	PR	prudentopolis	2
02-RJ-Porciuncula	110	113	6424618	.	RJ	porciuncula	2
02-RN-Barauna	108	65	3699752	108	RN	barauna	2
02-RO-Ouro_Preto_do_Oeste	140	100	2.68e+07	140	RO	ouro preto do oeste	2
02-RR-Amajari	74	88	2626137	74	RR	amajari	2
02-RS-Independencia	97	134	1536579	97	RS	independencia	2

Text as Data

- Classify an email message as either a legitimate email or spam
- Learn about the opinion of a politician on the topic of immigration
- The content of the text will certainly contain important information for the task
- Text data is usually represented as concatenation of characters. In any of the examples just given, the length of the text data will vary
- This feature is clearly very different from the numeric features, and we will need to process the data before we can apply algorithms to it

Preprocessing

- Make it useful for our purposes
- Simplify and lower dimensionality

Document - Term

$$X = \begin{pmatrix} 1 & 0 & 0 & \dots & 3 \\ 0 & 2 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 5 \end{pmatrix}$$

$X = N \times K$ matrix

- $N =$ Number of documents
- $K =$ Number of features

Preprocessing for Quantitative Text Analysis

Recipe for preprocessing: retain useful information

- Remove capitalization, punctuation
- Discard stop words
- Discard Word Order (Bag of Words Assumption)
- Create Equivalence Class: Stem, Lemmatize, or synonym
- Discard less useful features (depends on application)
- Other reduction, specialization

Output: Count vector, each element counts occurrence of stems

Stop Words

Stop Words: English Language place holding words

- the, it, if, a, able, at, be, because...
- Add “noise" to documents (without conveying much information)
- Discard stop words: focus on substantive words
- **Caution**: Exercise caution when discarding stop words. You may need to customize your stop word list.

Creating an Equivalence Class of Words

Reduce dimensionality further (create equivalence class between words)

- Words used to refer to same basic concept.
  - family, families, familial $\rightarrow$ famili
- Stemming/Lemmatizing algorithms: Many-to-one mapping from words to stem/lemma

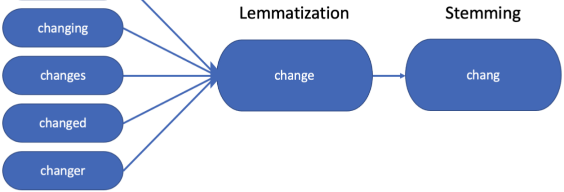
Stemming vs Lemmatization

- Stemming algorithm:
  - Consists of chopping off end of word
  - Porter stemmer, Lancaster stemmer, Snowball stemmer
- Lemmatizing algorithm:
  - Condition on part of speech (noun, verb, etc)
  - Verify result is a word

Stemming vs Lemmatization

- Stemming algorithm:
  - Word representations may not have any meaning
  - Takes less time
  - Use stemming when meaning of words is not important for analysis. Example: Spam detection.
- Lemmatizing algorithm:
  - Word representations have meaning
  - Takes more time than Stemming
  - Use lemmatization when meaning of words is important for analysis. Example: question answering application.

Stemming -vs- Lemmatization



Additional read

Stemming and Lemmatization – Stanford NLP
https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html

Preprocessing reduces dimensionality where it causes problems for inference (stopwords, stemming) and sometimes increases dimensionality when it makes our inferences better (bigram, ngrams)

A Complete Example

“Political power grows out of the barrel of a gun" - Mao

A Complete Example

“Political power grows out of the barrel of a gun" - Mao

- **ngram**: An analyst may want to combine words into a single term that can be analyzed.

A Complete Example

[Political], [power], [grows], [out], [of], [the], [barrel of a gun] - Mao

- **ngram**: An analyst may want to combine words into a single term that can be analyzed.

A Complete Example

[Political], [power], [grows], [out], [of], [the], [barrel of a gun] - Mao

- **Remove Stopwords**: Removing terms that do not convey important information

A Complete Example

[Political], [power], [grows], [out], [barrel of a gun] - Mao

- **Stemming**: Takes the ends of conjugated verbs or plural nouns, leaving just the stem.

A Complete Example

Finally, we can turn tokens and documents into a "document-term matrix."

- Imagine we have a second document in addition to the Mao quote, which tokenizes as follows

Document 1: [polit], [power], [grow], [out], [barrel gun]
Document 2: [compar], [polit], [chicago], [polit]

Transpose Document-Term-Matrix

$$
\begin{array}{c|cc}
 & Doc1 & Doc2 \\
power & 1 & 0 \\
grow & 1 & 0 \\
out & 1 & 0 \\
barrel\ of\ a\ gun & 1 & 0 \\
compar & 0 & 1 \\
polit & 1 & 2 \\
chicago & 0 & 1
\end{array}
$$

All steps together

1. Remove capitalization and punctuation
2. Discard word order (Bag of Words)
3. Remove stop words
4. Applying Stemming Algorithm
5. Create count vector

Vectorization a simple example

You have 2 documents:

1. Blue House
2. Red House

Our corpus will consist of all the words in the documents namely: Red, Blue, House. The vector representation in the “Bag of Words" approach:

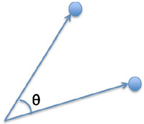
- "Blue House" -> (red, blue, house) -> (0, 1, 1)
- "Red House" -> (red, blue, house) -> (1, 0, 1)

Once we have vector representation, we can do analysis. Algorithms can handle numbers (vectors).

Cosine Similarity

Cosine similarity is a measure of similarity between two non-zero vectors of an inner product space. The cosine similarity is particularly used in positive space, text data, where the outcome is neatly bounded in $[0, 1]$.

$$sim(A, B) = cos(\Theta) = \frac{A \cdot B}{\|A\| \|B\|}$$



Term Frequency and Inverse Document Frequency

- Improve on Bag of Words by adjusting word counts based on their frequency in corpus (the group of all the documents)
- Use Term Frequency - Inverse Document Frequency (TF-IDF)

TF-IDF

TF-IDF term x in document y

$$TF_{x,y} \times log(\frac{N}{DF_x})$$

- $TF_{x,y}$ = frequency of x in y
- DF_x = number of documents containing x
- N total number of documents

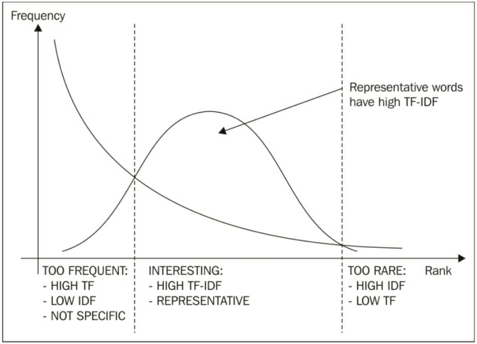
TF-IDF

TF Term- Frequency is the raw frequency of a word normalized by the number of words in the document

IDF Inverse Document Frequency is the number of documents normalized by the number of documents that contain the term. For terms that are present in every document, this will lead to an IDF value of zero (that is, log(1)). For this reason, one of the possible normalizations for IDF is $1+\log(N/DF_x)$.

TF-IDF

The intuition behind TF-IDF is that words which are too frequent or too rare are not representative



How can this work?

- Speech may contain sarcasm:
  - The Star Wars prequels were amazing because everyone loves a good discussion about trade policy
- Subtle Negation
  - They have not succeeded, and will never succeed, in breaking the will of this valiant people
- Order Dependence
  - Peace, no more war
  - War, no more peace

How Could This Possibly Work?

1. It might not: Validation is critical (task specific)
2. Central Tendency in Text: Words often imply what a text is about war, civil, union or tone consecrate, dead, died, lives. Likely to be used repeatedly: create a theme for an article
3. Human supervision: Inject human judgement (coders): helps methods identify subtle relationships between words and outcomes of interest

It is easier to capture some things than others