TEXT ANALYSIS

Christopher Gandrud

SG1022, City University London

- What is text analysis and why use it?
- Human vs. machine coding
- The general process
- Specific issues with machine coding
- Pros and Cons

"When we perform textual analysis on a text, we make an educated guess at some of the most likely interpretations that might be made of that text." (McKee 2003, 1)

WHY USE TEXT ANALYSIS?

You *use* and *contribute* to text analysis every day.

Google

kanye west is

kanye west is **dead**
kanye west is **jesus**
kanye west is **from**
kanye west is **worth**

(Some of you) are building a data set that will be used for text analysis right now.

People are *interacting more through texts* and so creating increasingly more (machine accessible) texts.

Massive new source of data for social science analysis.

We may have research questions where we conducted a survey with an open-ended question.

We need some systematic way to understand these texts and make comparisons across survey respondents.

We may have research questions where we want to interview a group of people that are hard to access, but who produce many texts.

For example, in an ideal world we may want to survey world leaders for their preferences on handling Syrian refugees. We may want to see how these preferences change over time.

World leaders don't give many interviews (especially not multiple interviews on the same topic over time), but they–often filtered through a press office–do create many texts.
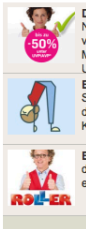
11. Januar 2016, 20:22 Uhr   Ermittlungen zu den Übergriffen in Köln

Kölner Polizei: Vor allem Marokkaner fallen auf

- Die Ermittler haben bisher 19 Tatverdächtige identifiziert.

- Neun Verdächtige halten sich illegal in Deutschland auf, zehn Personen sind Asylsuchende.

- Von den 19 Verdächtigen kommen 14 aus nordafrikanischen Ländern, vor allem aus Marokko.

- Den Statistiken der Kölner Ermittler zufolge werden 40 Prozent der nordafrikanischen Zuwanderer innerhalb eines Jahres straffällig.

Source: http://www.sueddeutsche.de/panorama/ermittlungen-zu-den-uebergriffen-in-koeln-vor-allem-marokkaner-fallen-auf-1.2814336

January 11, 2016, 20:22 Investigations on the attacks in Cologne

Cologne police: Especially Moroccans to fall

- Investigators have identified so far 19 suspects.

- Nine suspects keep illegal in Germany, ten people are asylum seekers.

- 14 Of the 19 suspects come from North African countries, mainly from Morocco.

- Statistics of Cologne investigators According to 40 percent of North African immigrants become delinquent within a year.

Source: http://www.sueddeutsche.de/panorama/ermittlungen-zu-den-uebergriffen-in-koeln-vor-allem-marokkaner-fallen-auf-1.2814336
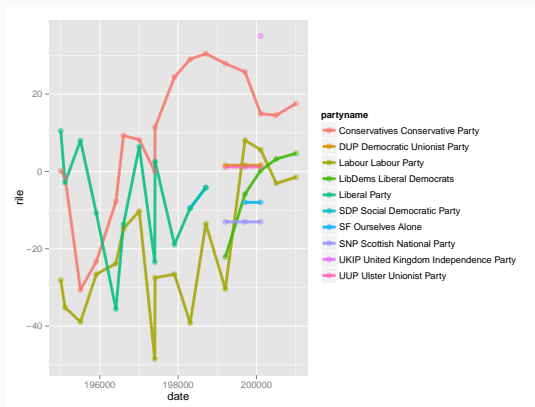
via Google Translate

We may have research questions about units that are not able to answer a questionnaire, but that produce texts.

E.g. International organisations, political parties, neighbourhood groups.

Left-Right Position of UK Parties Based on their Party Manifestos (*rile* negative values = more left-wing)
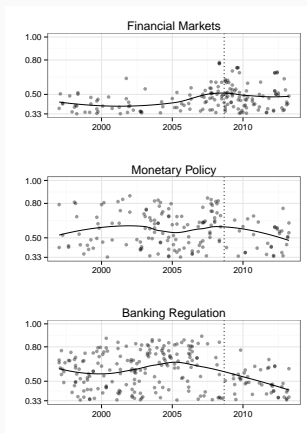
We may have research questions about how actors communicate to achieve goals.

For example, what topics do monetary policy bureaucrats talk about more when there is a financial crisis?

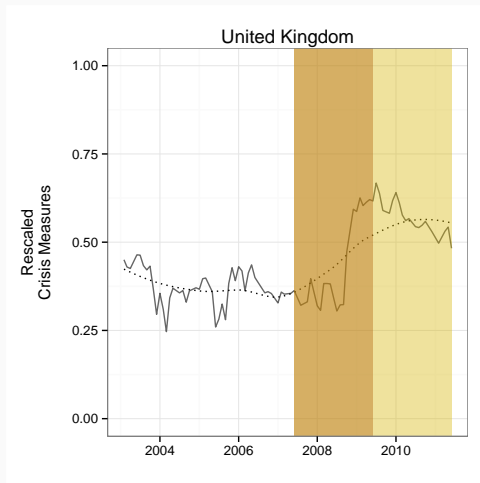(y-axis = proportion of speech discussing topic)

We may have research questions about widely held beliefs across time, for which a survey would be too costly or even impossible to run.

For example, if we wanted to study monthly perceptions of financial market stress across 180 countries.

(y-axis = level of financial stress)

(y-axis = level of financial stress)

Source: Gandrud and Hallerberg (2016)
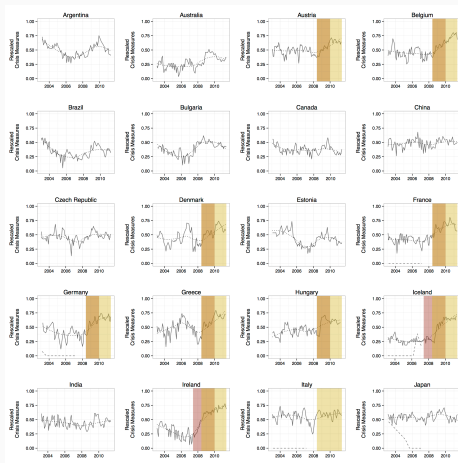
Or we want to study perceptions of world leaders over time…

(y-axis = larger values indicate more positive press)
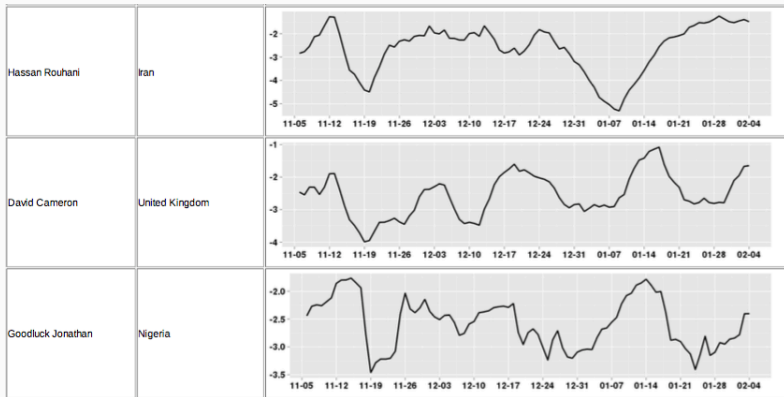
DEFINING TEXT ANALYSIS

"When we perform textual analysis on a text, we make an educated guess at some of the most likely interpretations that might be made of that text." (McKee 2003, 1)

"Content Analysis is a research technique for making replicable and valid inferences from texts (or other meaningful matter [e.g. videos, audio]) to the contexts of their use." (Krippendorff 2013, 24)

"Content Analysis is a research technique for making replicable and valid inferences from texts (or other meaningful matter [e.g. videos, audio]) to the contexts of their use."

Replicable: different researchers, independent of each other should get the same results when applying the same technique.

Replicable results are more reliable.

"Content Analysis is a research technique for making replicable and valid inferences from texts (or other meaningful matter [e.g. videos, audio]) to the contexts of their use."

Valid: research is open to careful scrutiny and your claims can be upheld given independently available evidence.

"Content Analysis is a research technique for making replicable and valid inferences from texts (or other meaningful matter [e.g. videos, audio]) to the contexts of their use."

Texts: something that is produced by someone to have meaning for someone else.

E.g. newspaper articles, treaties, transcripts, tweets, maps, advertisements, press releases, movies, party manifestos.

In this course we focus exclusively on written texts composed of words.

1. Texts have no objective–reader-independent qualities. Meaning (data) arises from someone reading the text, often expecting other's understanding.

1. Texts have no objective–reader-independent qualities. Meaning (data) arises from someone reading the text, often expecting other's understanding.
2. Texts do not have single meanings.

1. Texts have no objective–reader-independent qualities. Meaning (data) arises from someone reading the text, often expecting other's understanding.
2. Texts do not have single meanings.
3. Meanings invoked by texts need not be shared.

1. Texts have no objective–reader-independent qualities. Meaning (data) arises from someone reading the text, often expecting other's understanding.
2. Texts do not have single meanings.
3. Meanings invoked by texts need not be shared.

1. Texts have no objective–reader-independent qualities. Meaning (data) arises from someone reading the text, often expecting other's understanding.
2. Texts do not have single meanings.
3. Meanings invoked by texts need not be shared.
4. Texts have meanings relative to particular contexts.

1. Texts have no objective–reader-independent qualities. Meaning (data) arises from someone reading the text, often expecting other's understanding.

2. Texts do not have single meanings.

3. Meanings invoked by texts need not be shared.

4. Texts have meanings relative to particular contexts.

5. Content analysts infer answers to particular research questions from their texts. Their inferences are merely more systematic, explicitly informed, and verifiable …than what ordinary readers do.

Text analysis can create data that is:

- Discrete: e.g. the main topics of a text.

- Continuous: e.g. proportion of a document dedicated to a specific word or words, scale (negative to positive, left-right).

The choice largely depends on your research question.

HUMAN AND MACHINE CODING

You can analyse texts either by relying exclusively on human coders or also rely on machine-assistance.

Note: you should never exclusively rely on machine coding. At a minimum, you need to check the validity of your machine assigned codes.

Face validity: Do the machine assigned codes make sense in relation to the context?

Machine coding has the advantage of being much more efficient for large numbers of texts.

· For example, it would basically be impossible for GDELT to create a daily updated index of world leader press coverage with human coders.

Machine coding is often more easily reproducible and update-able.

Regardless of whether you use human or machine coding, the general text analysis process is the same.

GENERAL TEXT ANALYSIS STEPS

1. Define the population of texts you are interested in (e.g. press releases by a particular organisation, open-ended survey responses).

2. Gather your sample of texts

3. Develop a coding scheme and classify your texts.

4. Establish the reliability and validity of your classifications.

Modified from: http://psc.dss.ucdavis.edu/sommerb/sommerdemo/content/doing.htm

At least two items to consider when defining your population of texts:

· Should be relevant for your research question.

· Texts should be accessible.

As with all data gathering, how you sample your texts can greatly affect your results.

For example, if you want to code press attitudes towards immigrants, but only gather articles from *The Guardian*, you will get much different results than if you only sample *The Daily Mail*.

We will discuss sampling in more detail in Week 7.

New tools for automatically gathering–web scraping–large numbers of texts from the internet.

Can make the data collection process dramatically faster and more reproducible.

We do not cover these tools in this course.

In order to enhance reproducibility, when you gather your sample (your Corpus) it should be well-organised and electronically available.

Always consider reliability when developing your coding scheme.

· Will another coder make the same choices given only the information in your coding scheme?

So, always fully document your coding scheme and explain your rationale.

Determine if you want to create a discrete (e.g. main topic of the text) or continuous coding scheme (e.g. attitude scale).

This decision should be based on relevance to your research question.

Skim a sub-sample of the texts to make a list of possible topics or words that would indicate a particular attitude, etc.

From this initial list, create operational definitions of your topic categories or scale.

· In order to enable replication, make these definitions as clear and specific as possible.

· Check that your definitions are comprehensive. Do they cover as many topics, words related to attitudes as possible?

· Make sure that your definitions are mutually exclusive, i.e. there is no overlap.

Now, apply your coding scheme.

You should always have at least one other rater independently apply your coding scheme.

Then check the level of aggreement. Ideally, different coders will give the same codes to the same texts based on the same coding scheme.

· This is known as high inter-rater reliability
· Simple tests for this would be a correlation coefficient (for continuous codes) or X^2 tests (for discrete codes). Note: a lack of independence between the two raters' scores is what you want.

If there is considerable disagreement between raters, you need to re-evaluate your coding scheme and possibly recode your corpus.

If you used machine coding, then you should select a random sub-sample of the texts and check to see if the machine codes match your intended coding scheme.

SPECIAL ISSUES WITH MACHINE CODING

There are many different advanced techniques for machine coding:
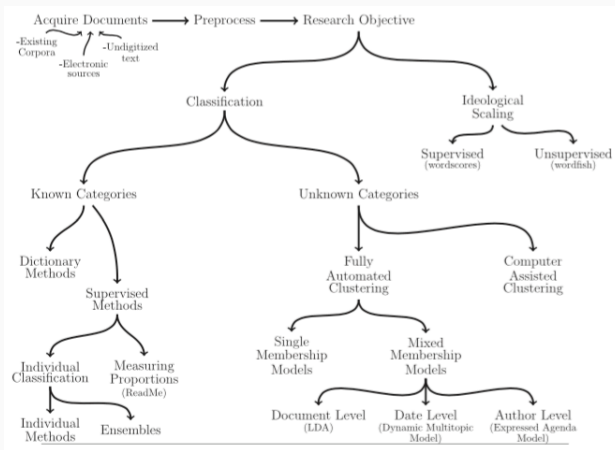
In this course we will focus on:

· The text preprocessing step.

· Simple word frequency methods of text analysis.

Regardless of the type of machine coding you use, you need to preprocess your texts.

This can include...

Removing unnecessary white space (spacing between words),
punctuation, capitalisation, numbers, etc.

Removing stopwords: function words that do not convey meaning like "a" and "the".

Stem your words: reduce the ends of words to reduce the total number of unique words.

- For example: *family*, *families*, *families'*, *familial*, are changed to their stem: *famili*.
- Stemming is related to linguistic concept called *lemmatization*.

Note: each preprocessing decision affects your results and so should be fully justified.

Once you have preprocessed your data, then you can have your computer code the corpus.

In this course, we are going to focus on word frequency methods.

Note, you should focus on the relative word frequency. Most simply:

$$\text{Relative Frequency} = \frac{\text{Freq. of word in text}}{\text{Total words in text}} \tag{1}$$

Corrects for words being more frequent because a text is longer.

Word frequency is an efficient way to help you reproducibly summarise many texts,

but...

word frequencies have no inherent meaning.

You still need to code what the frequency of a word means, relative to the context and your research question.

Refugees discussed in terms of the topic …

Refugees discussed in terms of the topic *crimminality*.

PROS AND CONS

Pros	Cons
Texts are a massive new source of social data	Results can be misleading if we don't appreciate the context within which the speech acts takes place.
Social interactions are increasingly happening via texts	Purely descriptive, need to do more work to understand why
Useful for tracking changes over time	Sampling bias (including if writers delete texts) can be a major challenge
Can be an inexpensive way to gather social data	