Introduction to Computational Content Analysis

Snorre Ralund, Ph.D Fellow, SoDaS, UCPH

KØBENHAVNS UNIVERSITET


BACK
TO THE OLDSKOOL

Content Analysis in the Social Sciences

"Language contain within it societal choices about how to represent and interrelate concepts. Such choices frame the wayu individuals think about the world, and so affact what actions individuals take. Thus language affects behavior"

Content Analysis in the Social Sciences

Content as Content

- Knowledge about what we are concerned with.
  - E.g. as a **Society** as seen in our **Political Attention**. Lasswell 1941: "The World Attention Survey"
  - Characterizing Cultures.

  English play less music in the radio than american, and are more occupied with the weather (Albig, William 1932)
  - As a **Scientific Field** i.e. a **Review**:

  "In spite of the all-too-obvious shortcomings of the method upon which the study is based, and defects in the study itself, the writer is of the opinion that some such approach must be used if "trends," "currents of thought,"" (Becker 1930)
- Knowledge about what we Know and When and how that changes.

Content of our interactions

Content as Content

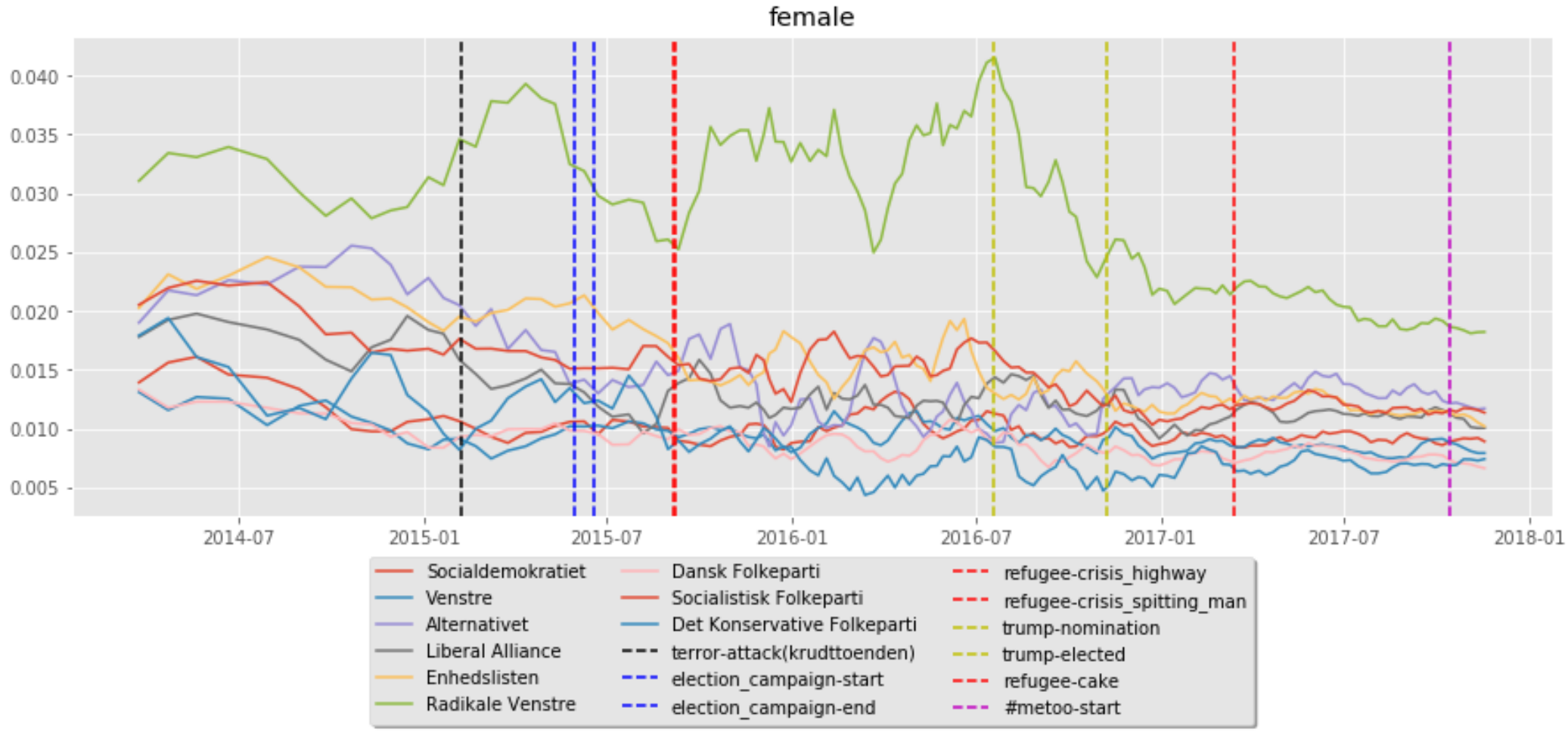
- Measuring historical development of jobdemand from jobpostings.
- Opportunism in politics:
  - Shifting political topics based on Response.
- Ethnic and Political Polarization through expressed social ties.
  - Processing the social signals in textual social media content.
  - Respect and Hostility towards the Stranger
- Expression and Reinforcement of group norms in social movements

Content of our interactions

Content as Content

- How are interactions between different ethnic groups changing in society?

Expressions of social relationships between Nordic Danes and Muslim Danes



Content Analysis in the Social Sciences

Content as Cause

- Communication content around us affects our behaviour and attitudes.
  - E.g. does movies cause crime? “Motion pictures and standards of Morality”
  - Are newspapers biased and are they biasing our opinions? “Newspaper bias in Congressional Controversies” (Kingbury and Heart 1933)

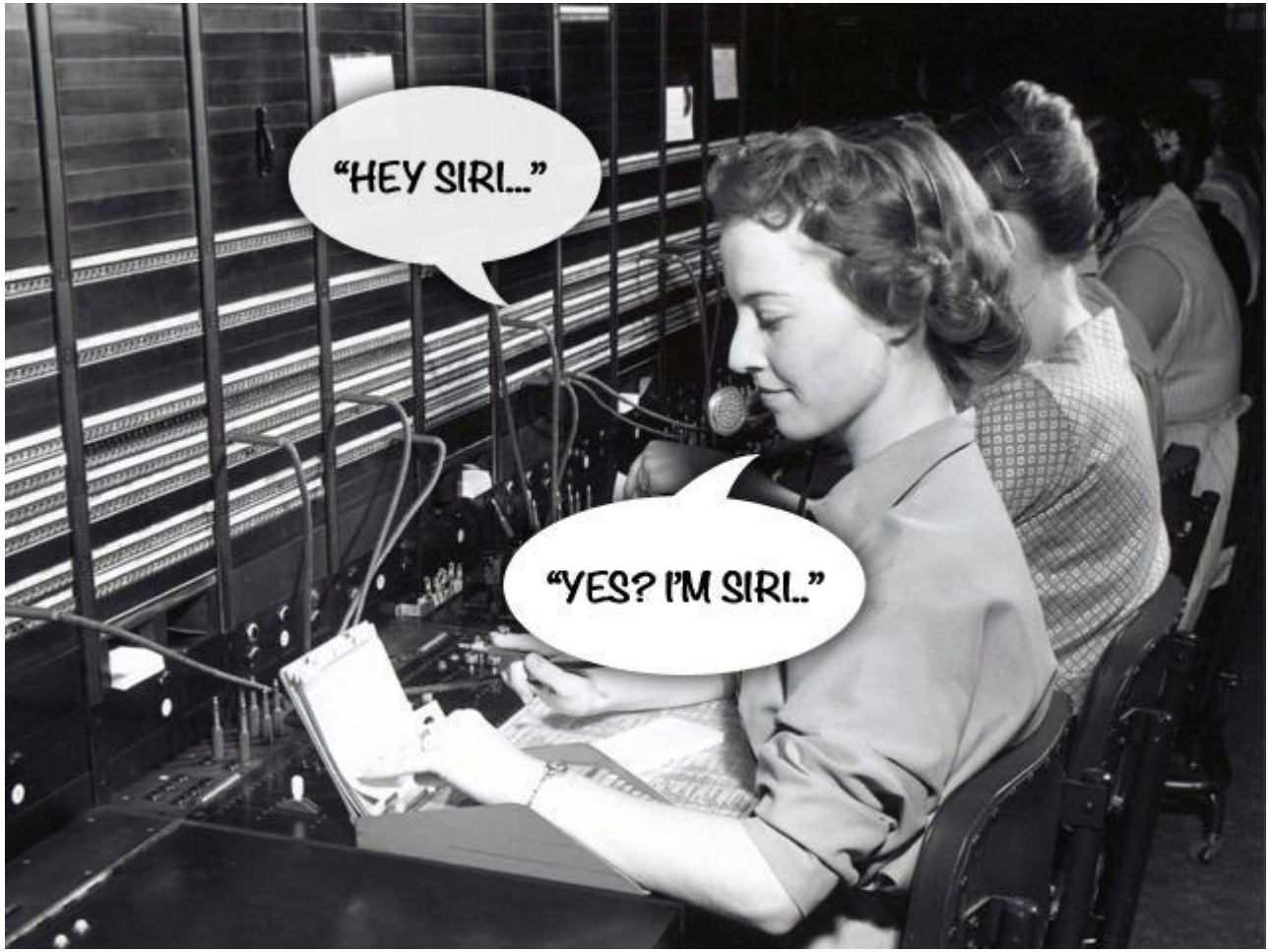

BACK
TO
THE FUTURE
OUTATIME

Natural Language Processing: State-of-the-Art (1)

Automatic speech recognition, CCG, Common sense, Constituency parsing, Coreference resolution, Dependency parsing, Dialogue, Domain adaptation, Entity linking, Grammatical error correction, Information extraction, Language modeling, Lexical normalization, Machine translation, Missing elements, Multi-task learning, Multi-modal, **Named entity recognition**, Natural language inference, **Part-of-speech tagging**, Question answering, Relation prediction, Relationship extraction, **Semantic textual similarity**, Semantic parsing, Semantic role labeling, **Sentiment analysis**, Shallow syntax, Simplification, Intent Detection and Slot Filling, Stance detection, Summarization, Taxonomy learning, Temporal processing, **Text classification**, Word sense disambiguation

Natural Language Processing: State-of-the-Art (1)

Speech to Text



Natural Language Processing: State-of-the-Art (1)

Human Reasoning

Natural Language Processing: State-of-the-Art (1)

Human Reasoning

Speech to Text: Google Cloud solution

Google Cloud solution:

SETUP

Install ggloud

```
pip install --upgrade google-cloud-speech
```

Get credentials

Follow guide to set up credentials. https://cloud.google.com/speech-to-text/docs/reference/libraries

enable datalogging for enhanced models

https://cloud.google.com/speech-to-text/docs/enable-data-logging

Speech to Text: Google Cloud solution

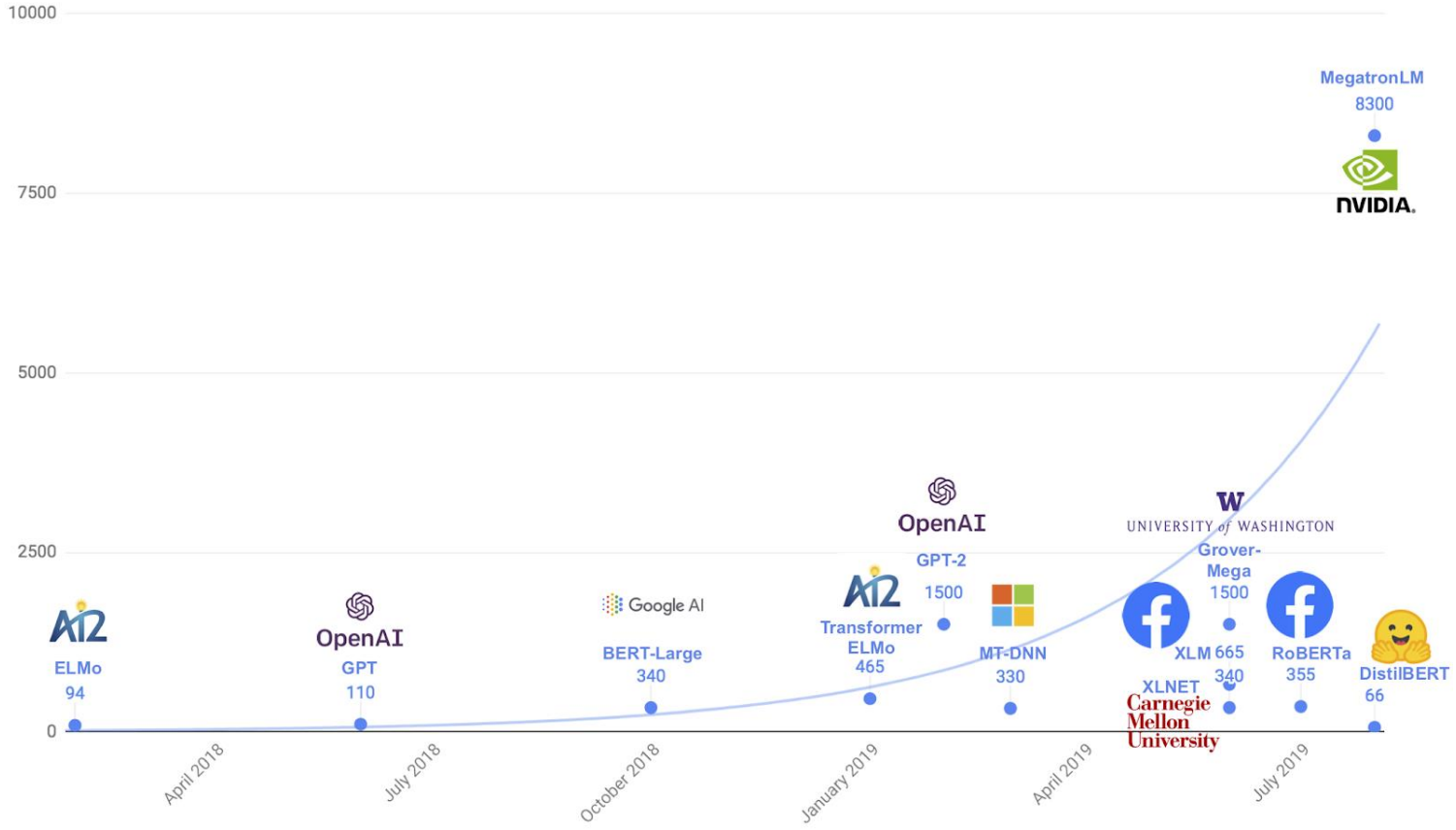
```
from google.cloud import speech_v1p1beta1 as speech2
credentials = '' # fill in the blank
client = speech2.SpeechClient.from_service_account_json(credentials)
fp = '' # path to a maximum 60 second audio file. See ffmpeg to divide file into parts programmatically
import io
with io.open(fp, 'rb') as audio_file:
    content = audio_file.read()
    audio = speech2.types.RecognitionAudio(content=content)

config = speech2.types.RecognitionConfig(
    audio_channel_count = n_channels,
    sample_rate_hertz=fs,
    language_code='da-DK',
    alternative_language_codes = ['ar-IQ'],
    max_alternatives = 5,
    profanity_filter=False,
    enable_word_time_offsets=True,
    enable_speaker_diarization=True,
    diarization_speaker_count=2,
    )
response = client.recognize(config, audio)
```

Natural Language Processing: State-of-the-Art (1)

Transfer Learning

Big models needs big data.

Models can be “recycled”!



Natural Language Processing: State-of-the-Art (1)

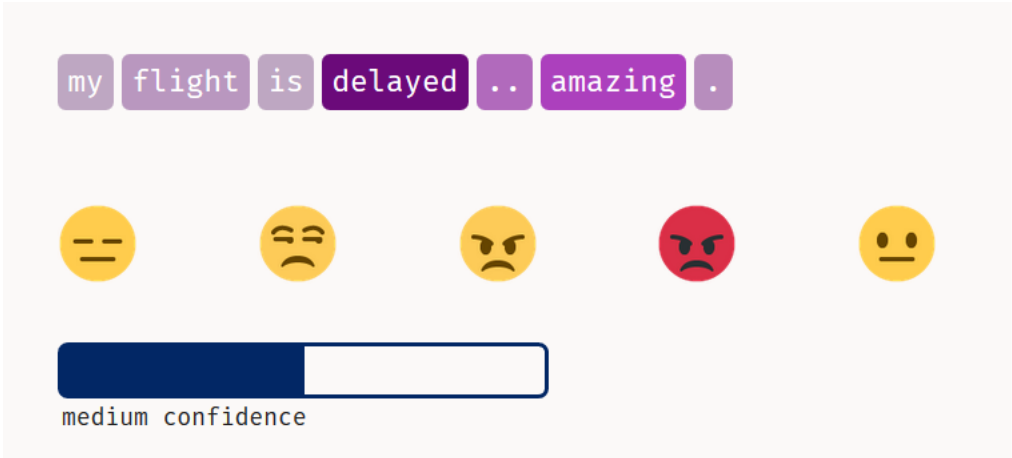
Noisy labels / Weak supervision

- **Deep Transfer Learning using noisy labels from emojies (Felbo et. al 2016)**
- **"Snorkel: rapid training data creation with weak supervision"(Ratner et. al 2019)**

Natural Language Processing: State-of-the-Art (1)

Sentiment Analysis

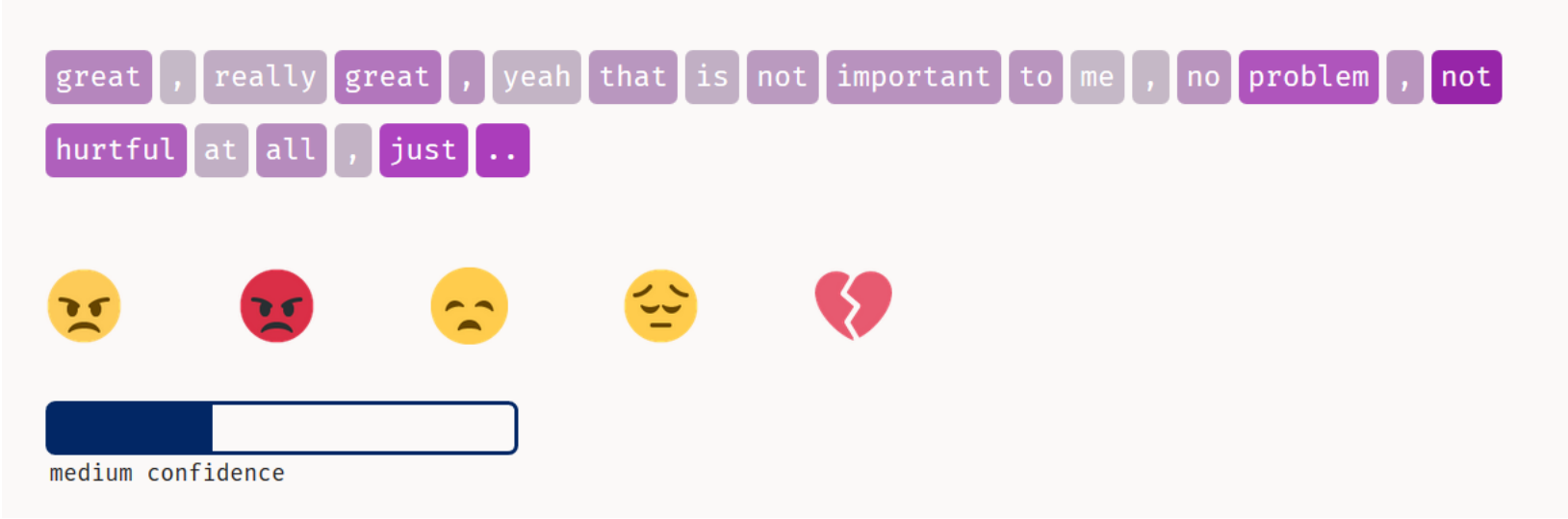
Sarcasm



Natural Language Processing: State-of-the-Art (1)

Sentiment Analysis

Sarcasm



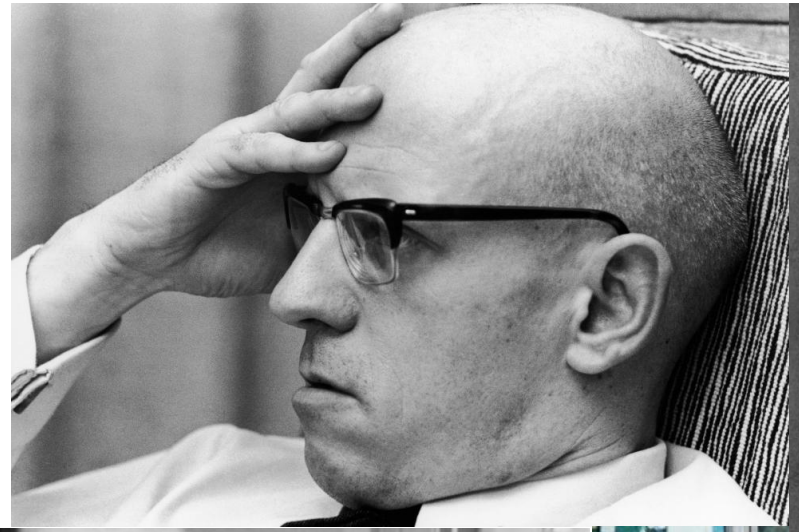
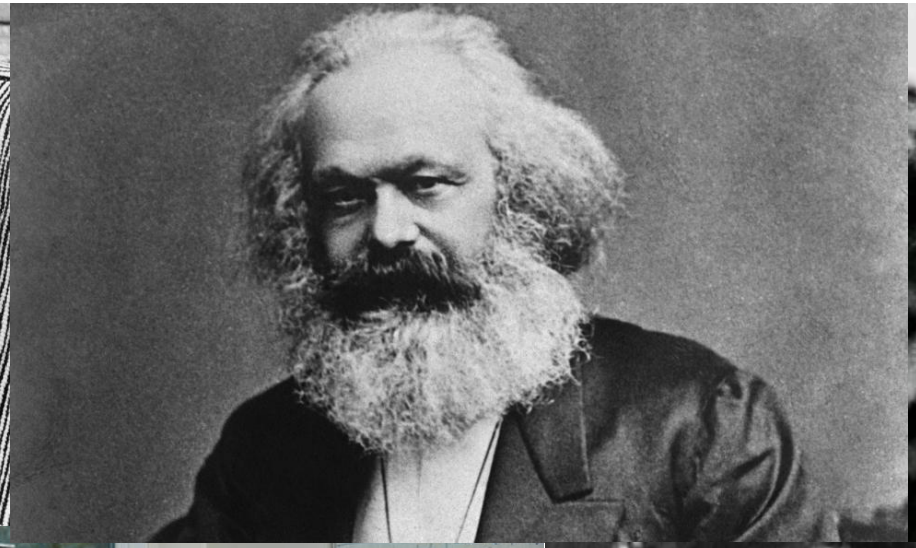
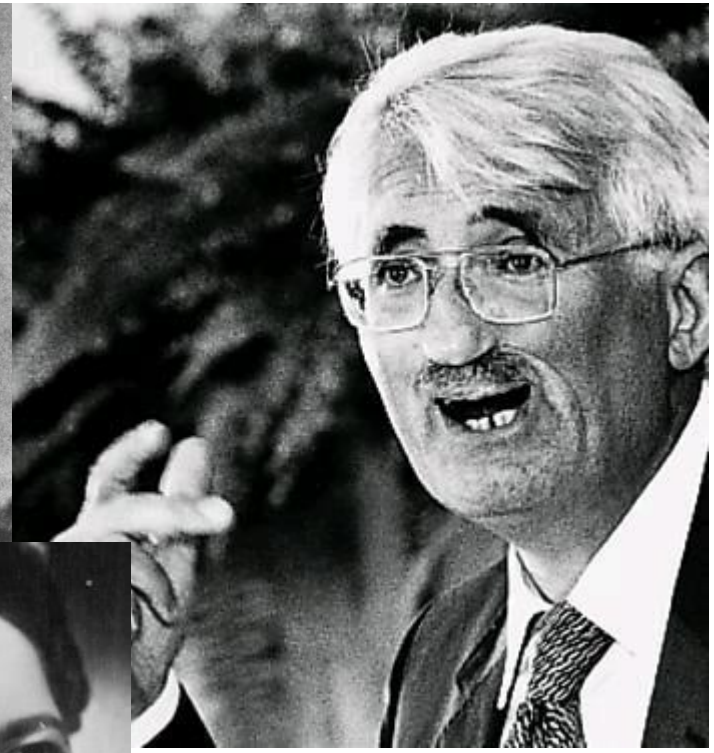
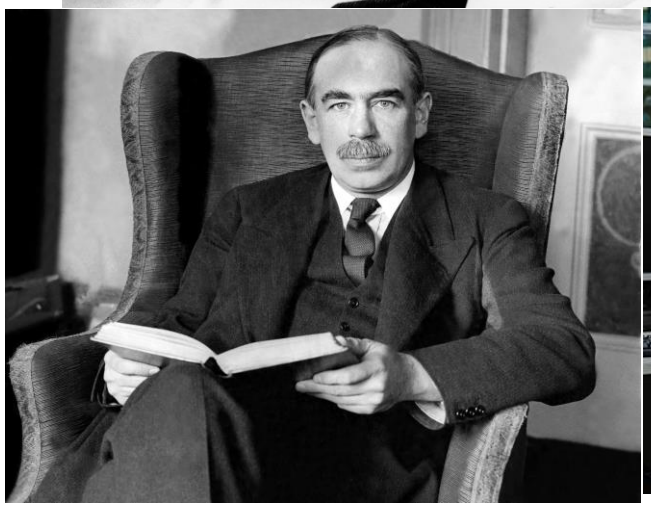
Potential Problems: Bias

```
Richard is muslim [(0.07356044, '🙏'), (0.05051396, '😠'), (0.048880223, '😕'), (0.044513963, '🙏'), (0.038552374, '♥')]
Richard is lesbian [(0.10796912, '😕'), (0.07081414, '😂'), (0.052409578, '😐'), (0.048823748, '💀'), (0.043848857, '😱')]
Richard is homosexual [(0.08500189, '😂'), (0.05664206, '💀'), (0.046665795, '😕'), (0.03654914, '😱'), (0.036210082, '😐')]
Richard is christian [(0.09523012, '😠'), (0.056955293, '🙏'), (0.050442047, '🙏'), (0.04362987, '😂'), (0.038723364, '💀')]
Richard is mormon [(0.07793487, '😂'), (0.049143367, '😕'), (0.044289604, '👏'), (0.040938176, '😠'), (0.0368746, '💀')]
Richard is gay [(0.07118924, '😐'), (0.06803788, '😑'), (0.06294611, '😂'), (0.05853562, '😒'), (0.05355592, '💀')]
```

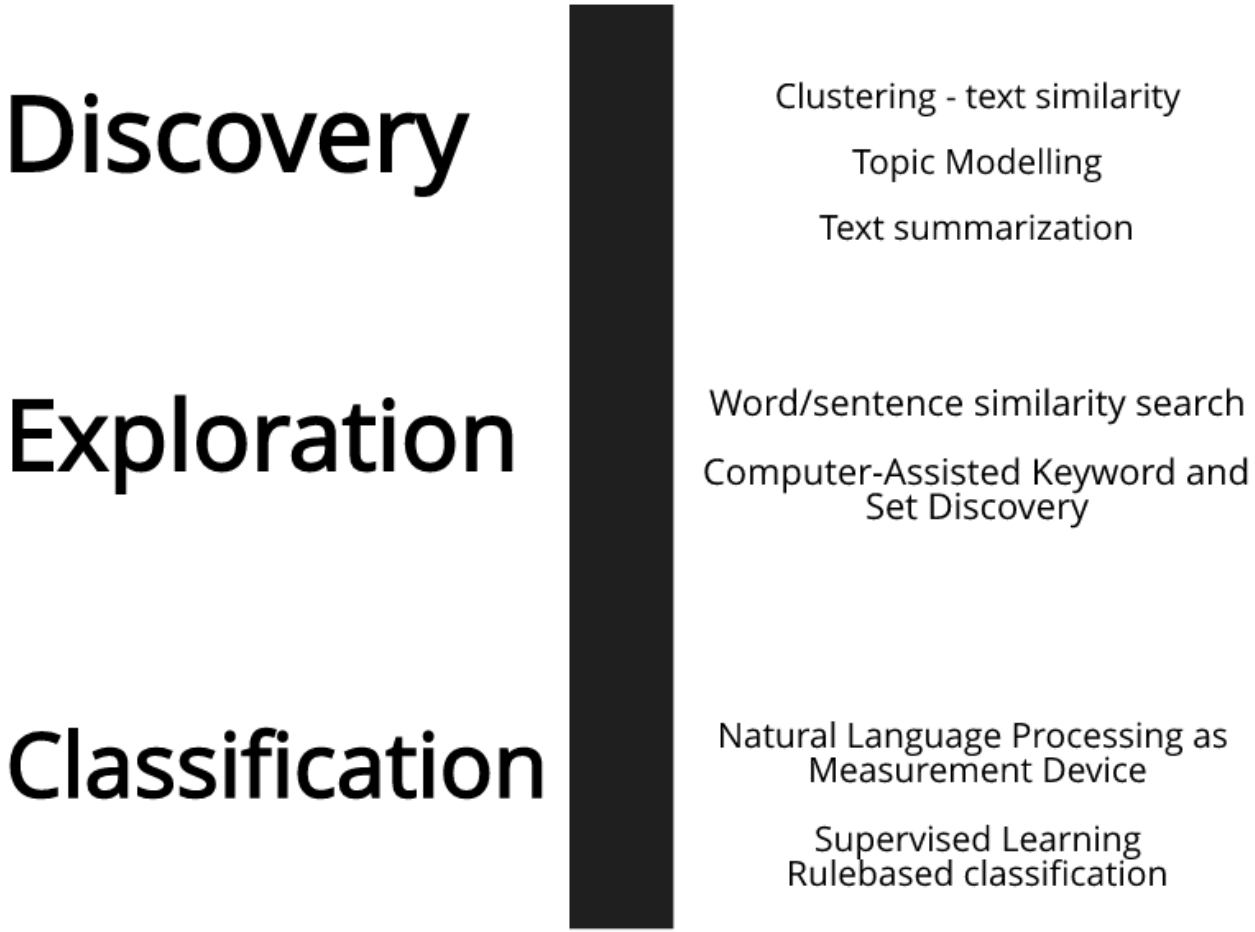
Potential Problems: Bias

Race	😠	😩	😭	😔	😳	😑	😕
African-American	0.026314	0.015215	0.022420	0.036655	0.039918	0.024493	0.03176
European	0.024837	0.012098	0.023012	0.037083	0.045626	0.024432	0.02759

NLP in the Social Sciences?

Computational Content Analysis



What do Social Scientist Do? (1)

Manual content analysis

- Manual reading, interpretation and systematic annotation of texts.
- **Lasswell 1941: "The World Attention Survey"**
- **Krippendorf 2018: Content Analysis: An Introduction to Its Methodology***

What do Social Scientist Do? (1)

Manual content analysis

- Manual reading, interpretation and systematic annotation of texts.

Pros

- Let humans do what they do best

BUT

- High costs cannot scale to big data
- Problems with training coders and in turn intercoder reliability.

What do Social Scientist Do? (2)

Dictionary methods and Word Counts

- **Sentiment Lexicons (LIWC), Topic Lexicions, Argument Lexicons etc.**

Examples

- Pennebaker and King 1999: "Linguistic styles: Language use as an individual difference."
- Graham, Haidt Nosek 2009: "Liberals and conservatives rely on different sets of moral foundations"
- Identification of suicidal behavior
- Menstrual Cycles inferred from autocorrelation in Emotional tweets.

What do Social Scientist Do? (2)

Dictionary methods and Word Counts

- **Sentiment Lexicons (LIWC), Topic Lexicions, Argument Lexicons etc.**

Pros

- Very efficient and simple to setup

BUT

- are they just random conglomerates with no consistent precision nor recall?
- Word ambivalences, and polysemi.
- Theoretical Validity? Is e.g. Negative emotions (i.e. sad and angry) actually a meaningful Category?

What do Social Scientist Do? (3)

Topic Modelling / Text clustering as measurement devices

Examples

- Grimmer 2012: *"Appropriators not position takers"*: The distorting effects of electoral incentives on congressional representation"
- Fligstein et al. 2017: *"The hegemony of a macro-economic framework in the Federal Open Market Committee during and before the financial crisis of 2008"*

What do Social Scientist Do? (3)

Topic Modelling / Text clustering as measurement devices

- Cheap and scalable(...)
- Datadriven and inductive categories
- Capture Polysemi (...).

BUT

- Arbitrary models and hyperparameter choices.
- Also needs explicit validation - i.e. Test Set - to prove its validity

What do Social Scientist Do? (3)

Topic Modelling / Text clustering as measurement devices

- Cheap and scalable(...)
- Datadriven and inductive categories
- Capture Polysemi (...).

BUT

- Arbitrary models and hyperparameter choices.
- Also needs explicit validation - i.e. Test Set - to prove its validity

What should social Data Scientist Do? (1)

State-of-the-art Deep Learning

- Learning from Bags-of-Words and atomized representations will not allow one to create high performing classifiers.

INSTEAD

- Use transfer learning (OpenAI, DeepMoji, ELMO, BERT).

Examples

- "Unsupervised Sentiment Neuron":https://openai.com/blog/unsupervised-sentiment-neuron/
- "DeepMoji": https://deepmoji.mit.edu/

What should social Data Scientist Do? (2)

Computationally grounded analysis: Datadriven grounding of categories.

- Social scientists cannot rely on categories / "theoretical schemata" developed ad hoc in other sciences. cf. Sentiment analysis.

INSTEAD

- We need a datadriven empirical understanding of what is important in the data.
- Using computational models (e.g. topic modelling) to effectively learn from large text data, and creating valid and empirically grounded categories.

What should social Data Scientist Do? (3)¶

Unbiased measurement and artificial effects

- When using supervised learning as a Measurement device beating state-of-the-art in classifier performance is not important to the social scientist.

INSTEAD

- We need to know and estimate the bias of our model, and correct it (see e.g. Hopkins & King 2010).
- We need to make sure that our models do not systematically underestimate certain categories (e.g. social class, gender, ethnicity etc).

What should social Data Scientist Do? (4)

Ethics and Conscience

Social data scientists appropriating these methods should:

- Be aware of the implications of your work
  - Who benefit from it, how can it be used and misused.
- Surveilance and privacy

Examples

- While Humans will not keep track of 28 days cycles in affectivity in each others social media content, statistics on emotional expressions tells us new things.
- Depression, ethnicity etc. detection used by Insurances.
- Creating a framework for detecting protest on chinese social media.