Investigation of Algorithms and Technologies used for Project

Investigation of Algorithms and Technologies used for Project

Thematic Analysis and opposing sentiment analysis based on news articles in python for natural language processing:

****

https://github.com/opinionated/OpinionatedNLPAnalysis (Note: articleDataExtraction does most of the heavy lifting). Uses NLTK, Watson Developer Cloud & AlchemyAPI)

****

Best Description of thematic Analysis:

https://www.interaction-design.org/literature/article/how-to-do-a-thematic-analysis-of-user-interviews

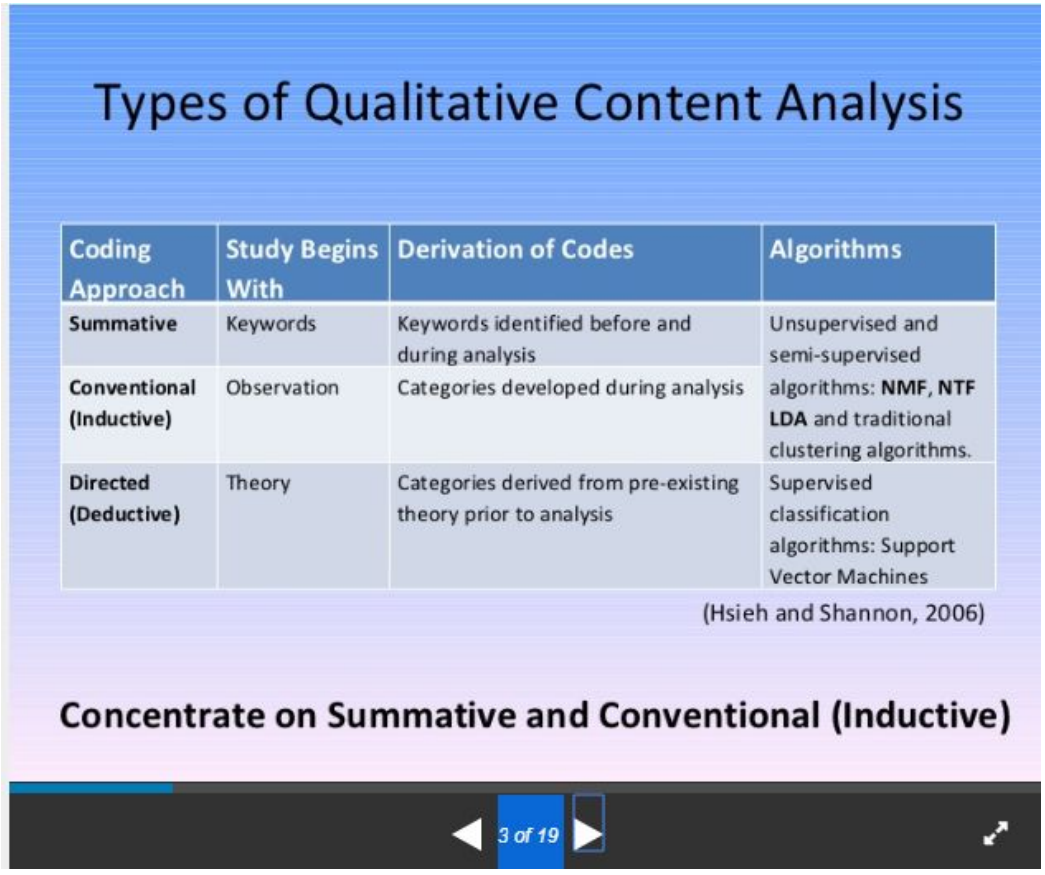

Image from:
https://www.slideshare.net/aneeshabakharia/algorithms-for-the-thematic-analysis-of-twitter-datasets2

Investigation of Algorithms and Technologies used for Project

Code & themes:

Need to pen and paper notes on text with examples to extend these catches.
Where to find open source examples of slips of the tongue?
-->http://mercercognitivepsychology.pbworks.com/w/page/33015683/Slip-of-the-Tongue%20Errors (Contains multiple examples)

E.g. if interested in streaming services -> HBO and netflix.
There is specific software for coding. (Research). Exploratory analysis finds themes on its own.

Examples for theme of slips of the tongue:

1. (I meant..)
2. (Sorry) -> Check word before and after and can tell if its a new word
3. Word (Followed by other word, but faster) -> Not sure how to include frequency in speech recognition. After some research, some hacky methods include catching where words are sped up drastically and assigning a certain non real word to them which signifies the word was sped up and it's a possible mistake.

Examples for theme of agreement:

1. Yes
2. Agree

After a great deal of research, the most optimal approach would not be thematic analysis at all for slips of the tongue, rather checking if the sentence is grammatically correct and if it's not then further boiling it down by checking if there is a word in the sentence that does not make sense, and then checking if the only thing that does not make sense is the misplacement of a real word, rather than stutter or some other mistake. Therefore, while thematic analysis may assist other finds such as signifiers (Homophones being the codes) it does not assist slips of the tongue.

Investigation of Algorithms and Technologies used for Project

Sentiment Analysis may have use to be monitored within the session

Perhaps gathering a score for anger, happiness, sadness etc. would be useful for the session

What are the use cases of WordNet and can it be applied?

How to use:

```
import nltk
nltk.download('wordnet')
```

WordNet can find if a word is a noun, verb, adjective and adverb.
Can find definitions, examples, Hypernyms - Similar words but related to each other like a tree.

NLP | Word Collocations seems that it has a lot of potential use, as it checks what words most commonly occur after a word, and can possibly be used to interpret meaning. (https://www.geeksforgeeks.org/nlp-word-collocations/)

nltk.probability.FreqDist can be used to count most used words. **ConditionalFreqDist** counts the most used tagged words for getting words that we want to find out.

Compare the similarity of words, for example, how similar are the words selling and hello? They are 27% similar as they share common hypernyms further up the two. Greeting and hello are 95% similar. https://www.geeksforgeeks.org/nlp-wupalmer-wordnet-similarity/.

Investigation of Algorithms and Technologies used for Project

Topic modeling using Latent Dirichlet Allocation (Build themes on transcripts) - Wont work (Not enough Data)

Will try to implement in movie scripts between therapist and psychologist (Sent in links beforehand and see if interesting topics come up).
(Really god implementation of this can be found in the document below)

- Generates topics based on word frequency in a set of documents..
- To implement LDA there is a requirement to:
    1. Gather sessions (From scripts, interviews etc.)
    2. Clean data:
        - Tokensize (Convert to atomic element).
        - Stopping (Removing meaningless words).
        - Stemming (Merge words that are equivalent in meaning).
- Count number of times each word appears
- Filter out words that appear less than X times in each document. Depends on data size. Genism library can do the filter.
- Create tf-idf (term frequency-inverse document frequency) model.
- Train LDA model using gensim.models.LdaMulticore
- For each topic explore words occuring in it and it's weight

Article explaining topic modeling implementation in python for news articles:

https://towardsdatascience.com/topic-modeling-and-latent-dirichlet-allocation-in-python-9bf156893c24
Source code:
https://github.com/susanli2016/NLP-with-Python/blob/master/LDA_news_headlines.ipynb

thematic analysis using NLTK

(No direct information on thematic analysis using NLTK).
https://github.com/opinionated/OpinionatedNLPAnalysis claims to use it, but didn't find it used directly inside the code, possibly implemented by the Alchemy API.

Investigation of Algorithms and Technologies used for Project

Alternative to NLTK

Tutorial setting up thematic analysis with SpaCey:

https://www.kaggle.com/caractacus/thematic-text-analysis-using-spacy-networkx
https://www.kaggle.com/gulsahdemiryurek/republican-and-democrat-tweets-analysis#4--Sentiment-Analysis (Sentiment)
spaCy i another natural language processing library that I built a basic script in to split nouns and verbs. Example script:

```
import spacy
import speech_to_text_speech_recognition as sr

# Load English tokenizer, tagger, parser, NER and word vectors
nlp = spacy.load("en_core_web_sm")

recording_name = 'harvard.wav'

# text = sr.transcribe_recording(recording_name)

text = 'the stale smell of old beer lingers it takes heat to bring'\
            'out the odour a cold dip restores health exist a salt
pickle'\
            'taste fine with him because of pasta are my favourite
exist for'\
            'food is the hot cross bun'

doc = nlp(text)

# Perform analysis on the text
print("Noun phrases:", [chunk.text for chunk in doc.noun_chunks])
print("Verbs:", [token.lemma_ for token in doc if token.pos_ ==
"VERB"])
```

NLTK can be used to split text into paragraphs.
(https://www.datacamp.com/community/tutorials/text-analytics-beginners-nltk)

Useful tutorial on NLTK:
https://medium.com/datadriveninvestor/python-data-science-getting-started-tutorial-nltk-2d8842fedfdd

Language processing book: https://nlp.stanford.edu/fsnlp/

Investigation of Algorithms and Technologies used for Project

Metaphone3

Metaphone3 allows the words from the recording to be translated into the phonetic representation of the word. This may be useful in being able to make a link of words that sound the same.

Good article on thematic analysis

http://tirap.ucalgaryblogs.ca/2016/07/21/how-to-conduct-thematic-analysis-for-sotl-research/