Text as Data: Computational Text Analysis

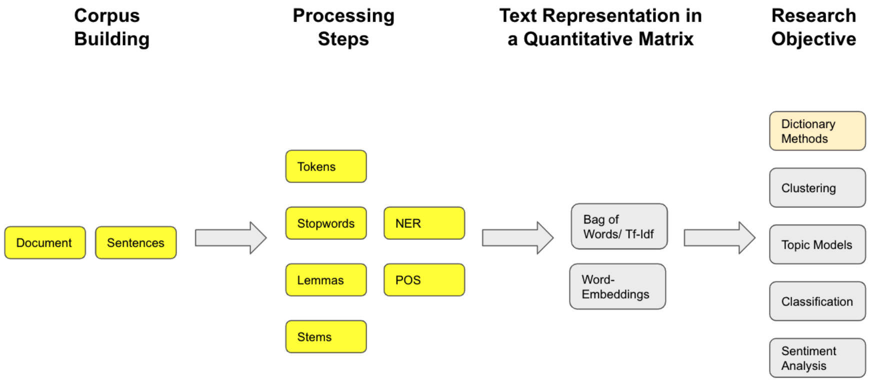
Week 2:
Natural Language Processing: Building Pre-Processing Pipeline, POS Tagging, NER + Building Dictionaries

Ashrakat Elshehawy

Department of Politics and International Relations,
University of Oxford

May 4, 2021

Overview from Text to Data[1]



[1] Some of the slides are based on Federico Nanni's course of Computational Text Analysis at U Mannheim

Recap - What have we learned until now?

- 4 Principles of Computational Text Analysis

Recap - What have we learned until now?

- 4 Principles of Computational Text Analysis
- Corpus Creation

Recap - What have we learned until now?

- 4 Principles of Computational Text Analysis
- Corpus Creation
- Text Pre-Processing Steps Matter

Recap - What have we learned until now?

- 4 Principles of Computational Text Analysis
- Corpus Creation
- Text Pre-Processing Steps Matter
- Stop-words

Recap - What have we learned until now?

- 4 Principles of Computational Text Analysis
- Corpus Creation
- Text Pre-Processing Steps Matter
- Stop-words
- Naive vs NLTK Tokenizer

Today

- Further on Text Pre-Processing (Stemming, Lemmatization)

Today

- Further on Text Pre-Processing (Stemming, Lemmatization)
- Pos-Tagging

Today

- Further on Text Pre-Processing (Stemming, Lemmatization)
- Pos-Tagging
- Named Entity Recognition

Today

- Further on Text Pre-Processing (Stemming, Lemmatization)
- Pos-Tagging
- Named Entity Recognition
- Building Dictionaries and Keyword Lists

Today

- Further on Text Pre-Processing (Stemming, Lemmatization)
- Pos-Tagging
- Named Entity Recognition
- Building Dictionaries and Keyword Lists
- The Power of Counting Words

Pre-Processing Example

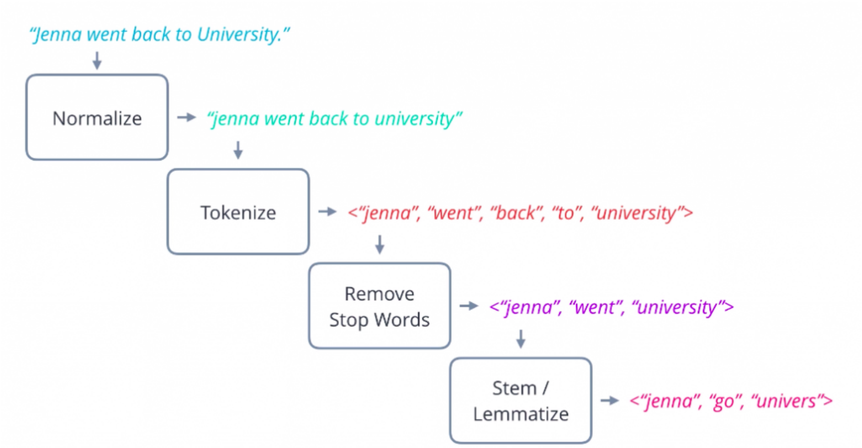


Image source: `https://eng.ftech.ai/`

Tokenization

Separating single words, starting from a string (which could be a document, a sentence, a tweet) $\rightarrow$ they will become tokens.

Tokenization

Separating single words, starting from a string (which could be a document, a sentence, a tweet) → they will become tokens.

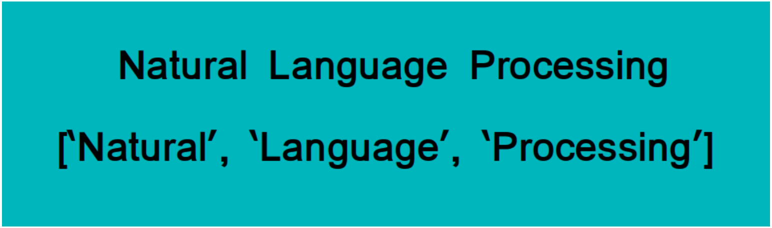


Image source: Analytics Vidha

Lemmatization

Lemmatization is the process of grouping together the inflected forms of a word as a single item. We "remove inflectional endings and to return the base or dictionary form of a word." (Schütze et al (2008) - NLP Stanford)

Lemmatization

Lemmatization is the process of grouping together the inflected forms of a word as a single item. We "remove inflectional endings and to return the base or dictionary form of a word." (Schütze et al (2008) - NLP Stanford)

- mice $\rightarrow$ mouse

Lemmatization

Lemmatization is the process of grouping together the inflected forms of a word as a single item. We “remove inflectional endings and to return the base or dictionary form of a word.” (Schütze et al (2008) - NLP Stanford)

- mice $\rightarrow$ mouse
- took $\rightarrow$ take

Lemmatization

Lemmatization is the process of grouping together the inflected forms of a word as a single item. We "remove inflectional endings and to return the base or dictionary form of a word." (Schütze et al (2008) - NLP Stanford)

- mice $\rightarrow$ mouse
- took $\rightarrow$ take
- studying $\rightarrow$ study
- fishes $\rightarrow$ fish

Stemming

Stemming is the process of reducing inflected words to their word stem. It refers to a “crude heuristic process that chops off the ends of words.” (Schütze et al (2008) - NLP Stanford)

Stemming

Stemming is the process of reducing inflected words to their word stem. It refers to a "crude heuristic process that chops off the ends of words." (Schütze et al (2008) - NLP Stanford)

- meeting $\rightarrow$ meet

Stemming

Stemming is the process of reducing inflected words to their word stem. It refers to a "crude heuristic process that chops off the ends of words." (Schütze et al (2008) - NLP Stanford)

- meeting $\rightarrow$ meet
- mice $\rightarrow$ mic

Stemming

Stemming is the process of reducing inflected words to their word stem. It refers to a "crude heuristic process that chops off the ends of words." (Schütze et al (2008) - NLP Stanford)

- meeting $\rightarrow$ meet
- mice $\rightarrow$ mic
- ponies $\rightarrow$ poni

Stemming

Stemming is the process of reducing inflected words to their word stem. It refers to a "crude heuristic process that chops off the ends of words." (Schütze et al (2008) - NLP Stanford)

- meeting → meet
- mice → mic
- ponies → poni

Sample text: Such an analysis can reveal features that are not easily visible from the variations in the individual genes and can lead to a picture of expression that is more biologically transparent and accessible to interpretation

Lovins stemmer: such an analys can reve featur that ar not eas vis from th vari in th individu gen and can lead to a pictur of expres that is mor biolog transpar and acces to interpres

Porter stemmer: such an analysi can reveal featur that ar not easili visibl from the variat in the individu gene and can lead to a pictur of express that is more biolog transpar and access to interpret

Paice stemmer: such an analys can rev feat that are not easy vis from the vary in the individ gen and can lead to a pict of express that is mor biolog transp and access to interpret

Image source: Schütze et al (2008)

Lemmatization and Stemming

The goal of both stemming and lemmatization is to reduce inflectional forms and sometimes derivationally related forms of a word to a common base form. - Schütze et al (2008)

Word-Sense-Disambiguation

WordNet is a lexical database, a dictionary of words online with a focus on the association of words.

Word-Sense-Disambiguation

WordNet is a lexical database, a dictionary of words online with a focus on the association of words.

Many semantic applications in Natural Language Processing benefit from such a large lexical database.

Word-Sense-Disambiguation

WordNet is a lexical database, a dictionary of words online with a focus on the association of words.

Many semantic applications in Natural Language Processing benefit from such a large lexical database.

With Word-Sense-Disambiguation we establish which sense of a word is used in a sentence, when a word is ambiguous and can have several meanings.

Word-Sense-Disambiguation

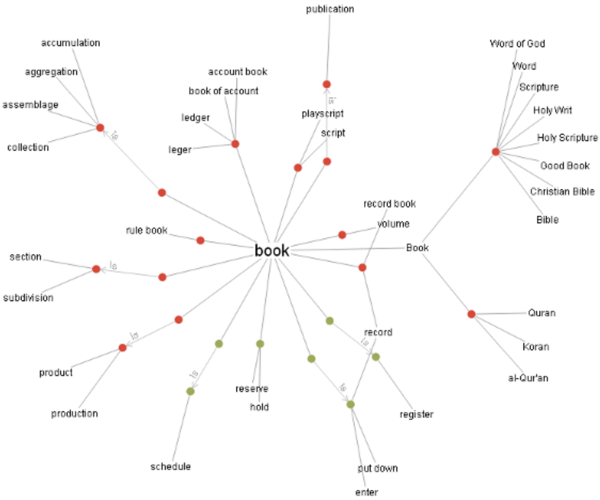


Image source: Open Science

Word-Sense-Disambiguation

Sentence 1: I can hear bass/frequency sound.
Sentence 2: He likes to eat grilled bass/fish

Employ WordNet and apply Lesk algorithm:

Word-Sense-Disambiguation

Sentence 1: I can hear bass/frequency sound.
Sentence 2: He likes to eat grilled bass/fish

Employ WordNet and apply Lesk algorithm:

- Retrieve all sense definitions of a target word

Word-Sense-Disambiguation

Sentence 1: I can hear bass/frequency sound.
Sentence 2: He likes to eat grilled bass/fish

Employ WordNet and apply Lesk algorithm:

- Retrieve all sense definitions of a target word
- Compare each sense definition with the sense definitions of the other words in context

Word-Sense-Disambiguation

Sentence 1: I can hear bass/frequency sound.
Sentence 2: He likes to eat grilled bass/fish

Employ WordNet and apply Lesk algorithm:

- Retrieve all sense definitions of a target word
- Compare each sense definition with the sense definitions of the other words in context
- Choose the sense with the highest overlap

Word-Sense-Disambiguation

Sentence 1: I can hear bass/frequency sound.
Sentence 2: He likes to eat grilled bass/fish

Employ WordNet and apply Lesk algorithm:

- Retrieve all sense definitions of a target word
- Compare each sense definition with the sense definitions of the other words in context
- Choose the sense with the highest overlap

We want to select the best sense for a word in a given context!

Word-Sense-Disambiguation

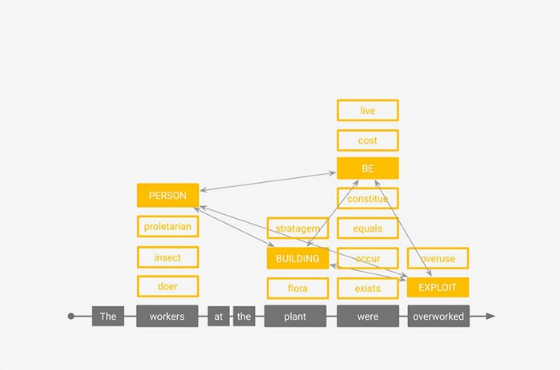


Image source: SAP Conversation AI 2015

Word-Sense-Disambiguation

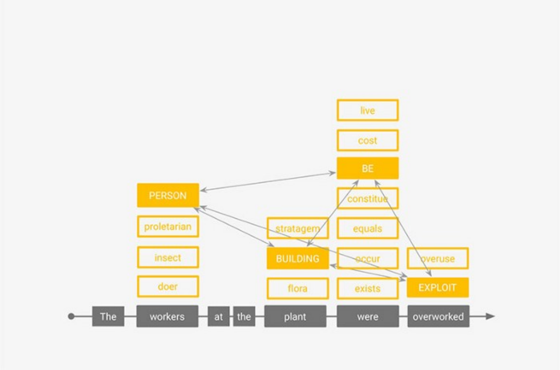


Image source: SAP Conversation AI 2015

- Assumption - words in a sentence should be part of a shared topic

Word-Sense-Disambiguation

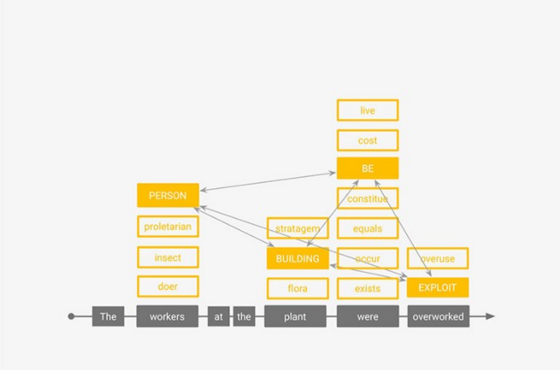


Image source: SAP Conversation AI 2015

- Assumption - words in a sentence should be part of a shared topic
- Algorithm 1. check a sentence, 2. "selects the senses whose definitions have the maximum overlap (the highest number of common words)" (SAP 2015)

Part-of-Speech-Tagging

The process of classifying words into their parts of speech and labeling them accordingly is known as part-of-speech (POS) tagging. - Schütze et al (2008)

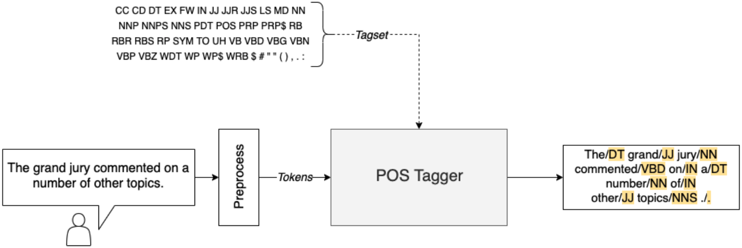


Image source: Devopedia 2019

Part-of-Speech-Tagging

One of the main problems are ambiguity.

They **refuse** to **permit** us to **obtain** the **refuse permit.**

Part-of-Speech-Tagging

They **refuse/VERB** to **permit** us to **obtain** the **refuse permit.**

Part-of-Speech-Tagging

They **refuse/VERB** to **permit/VERB** us to **obtain** the **refuse permit.**

Part-of-Speech-Tagging

They **refuse/VERB** to **permit/VERB** us to **obtain/VERB** the **refuse permit.**

Part-of-Speech-Tagging

They **refuse/VERB** to **permit/VERB** us to **obtain/VERB** the **refuse/NOUN permit/NOUN.**

Part-of-Speech-Tagging

Statistical POS tagging - we attach descriptive tags to each token to show which parts of speech are these tokens associated to. Tags are usually: verbs, noun, adjectives, etc.

Time/Noun flies like an arrow

- Capital: Yes
- Length: 4
- Prefix: No
- Suffix: No
- Beginning of Sentence: Yes

Part-of-Speech-Tagging

Statistical POS tagging - we attach descriptive tags to each token to show which parts of speech are these tokens associated to. Tags are usually: verbs, noun, adjectives, etc.

Time/Noun flies/Verb? OR noun? like an arrow

Part-of-Speech-Tagging

Statistical POS tagging - we attach descriptive tags to each token to show which parts of speech are these tokens associated to. Tags are usually: verbs, noun, adjectives, etc.

Time/Noun flies /Verb? OR noun? like an arrow

Part-of-Speech-Tagging

Statistical POS tagging - we attach descriptive tags to each token to show which parts of speech are these tokens associated to. Tags are usually: verbs, noun, adjectives, etc.

Time/Noun flies /Verb? OR noun? like an arrow

- Probability of *flies* being a verb or a noun

Part-of-Speech-Tagging

Statistical POS tagging - we attach descriptive tags to each token to show which parts of speech are these tokens associated to. Tags are usually: verbs, noun, adjectives, etc.

Time/Noun flies /Verb? OR noun? like an arrow

- Probability of *flies* being a verb or a noun
- Probability of a verb (or a noun) following a noun

Part-of-Speech-Tagging

Statistical POS tagging - we attach descriptive tags to each token to show which parts of speech are these tokens associated to. Tags are usually: verbs, noun, adjectives, etc.

Time/Noun flies /Verb? OR noun? like an arrow

- Probability of *flies* being a verb or a noun
- Probability of a verb (or a noun) following a noun

→ Hidden Markov Model - using both tag sequence probabilities and word frequency measurements.

Named Entity Recognition

NER seeks to locate and classify pieces of text into predefined categories such as the names of:

- persons
- organizations
- locations
- expressions of times
- quantities
- monetary values
- percentages

Monday, October 30, Hillary Clinton will present her book in Chicago at the University of Chicago.

Named Entity Recognition

NER seeks to locate and classify pieces of text into predefined categories such as the names of:

- persons
- organizations
- locations
- expressions of times
- quantities
- monetary values
- percentages

Monday, October 30, Hillary Clinton will present her book in Chicago at the University of Chicago.

Named Entity Recognition

First of all, regular expression to extract:

- telephone numbers
- E-mails
- Dates
- Prices
- Locations (e.g., word + "river" indicates a river $\rightarrow$ Hudson river)

Named Entity Recognition

Then, gazetteers with list of proper names of:

- Person
- Location
- Organization

Named Entity Recognition

Then, context patterns, such as:

- PERSON earns [Money]
- PERSON joined [ORGANIZATION]
- PERSON fly to [LOCATION]

Dictionaries and Key-word Lists

Counting frequencies of words and dictionary methods are some of the most basic computational text analysis tools.

Dictionaries and Key-word Lists

Counting frequencies of words and dictionary methods are some of the most basic computational text analysis tools.

Examples why dictionaries can be helpful:

Dictionaries and Key-word Lists

Counting frequencies of words and dictionary methods are some of the most basic computational text analysis tools.

Examples why dictionaries can be helpful:

- Create your own dictionaries of terms you want to examine in a certain dataset (e.g Political Speeches) $\rightarrow$ Time-intensive task

Dictionaries and Key-word Lists

Counting frequencies of words and dictionary methods are some of the most basic computational text analysis tools.

Examples why dictionaries can be helpful:

- Create your own dictionaries of terms you want to examine in a certain dataset (e.g Political Speeches) $\rightarrow$ Time-intensive task
- Sentiment Analysis

Dictionaries and Key-word Lists

Counting frequencies of words and dictionary methods are some of the most basic computational text analysis tools.

Examples why dictionaries can be helpful:

- Create your own dictionaries of terms you want to examine in a certain dataset (e.g Political Speeches) $\rightarrow$ Time-intensive task
- Sentiment Analysis

Remember to always validate!

Dictionaries and Key-word Lists

Counting frequencies of words and dictionary methods are some of the most basic computational text analysis tools.

Examples why dictionaries can be helpful:

- Create your own dictionaries of terms you want to examine in a certain dataset (e.g Political Speeches) $\rightarrow$ Time-intensive task
- Sentiment Analysis

Remember to always validate! Examples: Manual coding, dropping words from the dictionary, etc.

Dictionaries and Key-word Lists - Example

Rooduijn, M. and Pauwels, T., 2011. Measuring populism: Comparing two methods of content analysis. West European Politics, 34(6)

- Research Question: How to measure populism?

Dictionaries and Key-word Lists - Example

Rooduijn, M. and Pauwels, T., 2011. Measuring populism: Comparing two methods of content analysis. West European Politics, 34(6)

- Research Question: How to measure populism?
- Problem Definition: Populism definition based on two components: people-centrism and anti-elitism

Dictionaries and Key-word Lists - Example

Rooduijn, M. and Pauwels, T., 2011. Measuring populism: Comparing two methods of content analysis. West European Politics, 34(6)

- Research Question: How to measure populism?
- Problem Definition: Populism definition based on two components: people-centrism and anti-elitism
- Data: Unit of Analysis are Election Manifestos of parties in Western Europe

Dictionaries and Key-word Lists - Example

Rooduijn, M. and Pauwels, T., 2011. Measuring populism: Comparing two methods of content analysis. West European Politics, 34(6)

- Research Question: How to measure populism?
- Problem Definition: Populism definition based on two components: people-centrism and anti-elitism
- Data: Unit of Analysis are Election Manifestos of parties in Western Europe
- Method: Classical Content Analysis with Coders vs. Dictionary based Computational Text Analysis with a pre-defined dictionary → Counts of words related to populism

Dictionaries and Key-word Lists - Example

Rooduijn, M. and Pauwels, T., 2011. Measuring populism: Comparing two methods of content analysis. West European Politics, 34(6)

- Research Question: How to measure populism?
- Problem Definition: Populism definition based on two components: people-centrism and anti-elitism
- Data: Unit of Analysis are Election Manifestos of parties in Western Europe
- Method: Classical Content Analysis with Coders vs. Dictionary based Computational Text Analysis with a pre-defined dictionary $\rightarrow$ Counts of words related to populism
- Building the dictionary: Based on theoretical and empirical reasoning (anti-elitism)

Dictionaries and Key-word Lists - Example

Rooduijn, M. and Pauwels, T., 2011. Measuring populism: Comparing two methods of content analysis. West European Politics, 34(6)

- Research Question: How to measure populism?
- Problem Definition: Populism definition based on two components: people-centrism and anti-elitism
- Data: Unit of Analysis are Election Manifestos of parties in Western Europe
- Method: Classical Content Analysis with Coders vs. Dictionary based Computational Text Analysis with a pre-defined dictionary → Counts of words related to populism
- Building the dictionary: Based on theoretical and empirical reasoning (anti-elitism)
- Validation:

Dictionaries and Key-word Lists - Example

Rooduijn, M. and Pauwels, T., 2011. Measuring populism: Comparing two methods of content analysis. West European Politics, 34(6)

- Research Question: How to measure populism?
- Problem Definition: Populism definition based on two components: people-centrism and anti-elitism
- Data: Unit of Analysis are Election Manifestos of parties in Western Europe
- Method: Classical Content Analysis with Coders vs. Dictionary based Computational Text Analysis with a pre-defined dictionary $\rightarrow$ Counts of words related to populism
- Building the dictionary: Based on theoretical and empirical reasoning (anti-elitism)
- Validation:
  - Face validity (Are populist parties populist?)

Dictionaries and Key-word Lists - Example

Rooduijn, M. and Pauwels, T., 2011. Measuring populism: Comparing two methods of content analysis. West European Politics, 34(6)

- Research Question: How to measure populism?
- Problem Definition: Populism definition based on two components: people-centrism and anti-elitism
- Data: Unit of Analysis are Election Manifestos of parties in Western Europe
- Method: Classical Content Analysis with Coders vs. Dictionary based Computational Text Analysis with a pre-defined dictionary → Counts of words related to populism
- Building the dictionary: Based on theoretical and empirical reasoning (anti-elitism)
- Validation:
  - Face validity (Are populist parties populist?)
  - Concurrent validity (compare both methods)

Dictionaries and Key-word Lists - Example

Rooduijn, M. and Pauwels, T., 2011. Measuring populism: Comparing two methods of content analysis. West European Politics, 34(6)

- Research Question: How to measure populism?
- Problem Definition: Populism definition based on two components: people-centrism and anti-elitism
- Data: Unit of Analysis are Election Manifestos of parties in Western Europe
- Method: Classical Content Analysis with Coders vs. Dictionary based Computational Text Analysis with a pre-defined dictionary $\rightarrow$ Counts of words related to populism
- Building the dictionary: Based on theoretical and empirical reasoning (anti-elitism)
- Validation:
  - Face validity (Are populist parties populist?)
  - Concurrent validity (compare both methods)
  - Reliability (split-half test)

The Power of Counting

- Compare word frequencies between groups/individuals (e.g., most common words per speaker in a debate).

The Power of Counting

- Compare word frequencies between groups/individuals (e.g., most common words per speaker in a debate).
- Compare corpora and find words distinctive to each

The Power of Counting

- Compare word frequencies between groups/individuals (e.g., most common words per speaker in a debate).
- Compare corpora and find words distinctive to each
- Remove infrequently used terms to improve performance (Denny and Spirling 2018)

The Power of Counting

- Compare word frequencies between groups/individuals (e.g., most common words per speaker in a debate).
- Compare corpora and find words distinctive to each
- Remove infrequently used terms to improve performance (Denny and Spirling 2018)
- Sentiment Analysis using word counts of pre-defined dictionaries (Young and Soroka 2012)

The Power of Counting

Nielsen, Richard. 2019. "What Counting Words Can Teach Us About Middle East Politics" APSA MENA Newsletter Volume 2, Issue 2, Fall 2019.

Question: "Why do Salafi women cite the hadith and Quran only half as often as men when they write online?" (Nielsen 2019)

- Male Preachers use "God" more than women

Table 1 Men Use Hadith-Related Phrases More Than Women

	% of Documents Using Term		% of All Words	
Term	Men	Women	Men	Women
allāh	95	85	3.02	1.60
ṣalā allāh	67	32	0.50	0.21
raḍī allāh	50	20	0.18	0.10
ḥadīth	64	39	0.30	0.14
qāl/yaqūl (al-)rasūl	30	10	0.047	0.026
Ibn Taymiyya	23	3	0.030	0.006
al-Bukhari	41	10	0.098	0.030
al-Bayhaqi	17	1	0.013	0.001
Abu Hurayra	29	4	0.048	0.008
al-Tabarani	16	1	0.014	0.002
Abu Dawud	29	4	0.043	0.006

Note: This table counts the use of hadith-related phrases in 3,470 documents (1,306,641 words) by women and 17,854 documents (74,991,711 words) by men on saaid.net. The left two columns show the percentage of men's and women's documents that contain each phrase. The right two columns show the percentage of men's and women's words devoted to each phrase. All gender differences are statistically significant at the .05 percent level.

Image source: Nielsen (2020)

References 1/2

- Denny, M. J. and Spirling, A., 2018. Text Preprocessing For Unsupervised Learning: Why It Matters, When It Misleads, And What To Do About It. Political Analysis 26(2), pp. 168–189
- Nielsen, Richard. 2019. "What Counting Words Can Teach Us About Middle East Politics" APSA MENA Newsletter Volume 2, Issue 2, Fall 2019.
- Nielsen, R.A., 2020. Women's Authority in Patriarchal Social Movements: The Case of Female Salafi Preachers. American Journal of Political Science, 64(1), pp.52-66
- SAP Conversation AI, "From context to user understanding", last modified Nov, 20 2015 at Medium `https://medium.com/@SAPCAI/from-context-to-user-understanding-a692b11d95aa`

References 2/2

- Schrodt, P.A., 2001, February. Automated Coding of International Event Data using Sparse Parsing Techniques. In Annual Meeting of the International Studies Association, Chicago.
- Schütze, H., Manning, C.D. and Raghavan, P., 2008. Introduction to Information Retrieval (Vol. 39, pp. 234-265). Cambridge: Cambridge University Press.
- Rooduijn, M. and Pauwels, T., 2011. Measuring populism: Comparing two methods of content analysis. West European Politics, 34(6), pp.1272-1283.
- Young, L. and Soroka, S., 2012. Affective News: The Automated Coding of Sentiment in Political Texts. Political Communication, 29(2), pp.205-231.