Classification and Categorization

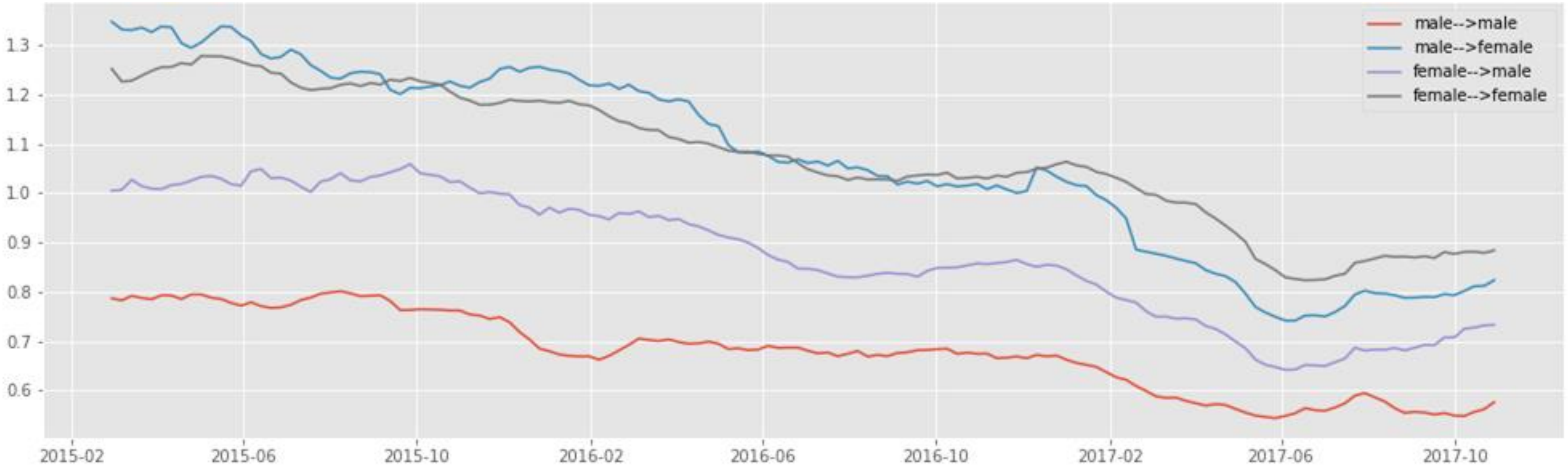
Computer Assisted Category Development

Snorre Ralund, Ph.D Fellow, SoDaS, UCPH

KØBENHAVNS UNIVERSITET

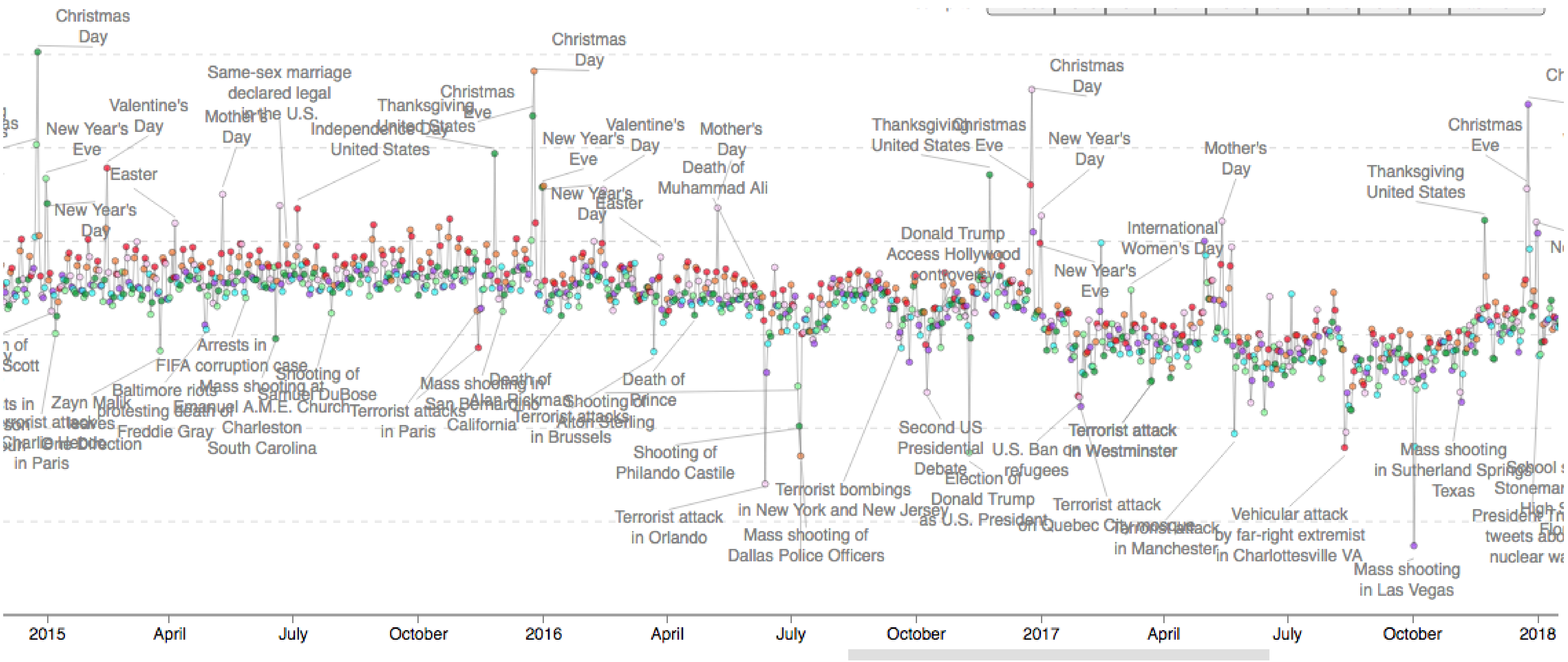
"Before we can trust the machine doing the interpretation, you need to demonstrate that you understand the data and the categorization"

Example: Online hostility between social groups



What is the Context?

What is relevant?



Some definitions

- CCA Validity:
  - *"A measuring instrument is considered valid if it measures what the user claims it measures."*(Krippendorf 2004: 313)

Some definitions

- Semantic Validity:
  - *"Semantic validity is the degree to which the analytical categories of texts correspond to the meanings these texts have for particular readers" (Krippendorf 2004: 323)*
  - E.g. Conglomerate Categories: Sentiment Analysis: Positive / Negative. What is a Funeral, what is Aggression,
- "Validity" as delegation of reading

  If you where to read this text: then you would say that it was about X.
  - This means that a-b has to agree on the definition of X (or A should define it well).
  - X should be clearly defined.
  - X should be clearly communicated.

Some definitions

Semantic Validity:

- *"The degree to which analytical categories accurately describe meanings and uses in the chosen contexts."*

- Distinction between the Readers *interpretation* and the Authors *intention.*
  - This has to do with understanding the context.

Reliability

- Validity means that we know what we are measuring, and that this is meaningful. Semantic Correspondance between the Category Definition and the Content Labeled.
- Reliability we allow for the instrument to not be 100 % efficient, that errors occur, and ask how well is the instrument working?

Reliability

- Validity means that we know what we are measuring, and that this is meaningful. Semantic Correspondance between the Category Definition and the Content Labeled.
- Reliability we allow for the instrument to not be 100 % efficient, that errors occur, and ask how well is the instrument working?
- Human coders are not 100 % reliable, this presumable reflects some "randomness" in the form of missing concentration similar to lazy respondents of a questionaire. But more importantly it could reflect inherent problems and ambivalences in the category definitions / coding scheme.

Reliability

- Validity means that we know what we are measuring, and that this is meaningful. Semantic Correspondance between the Category Definition and the Content Labeled.
- Reliability we allow for the instrument to not be 100 % efficient, that errors occur, and ask how well is the instrument working?

$$\alpha = 1 - \frac{D_o}{D_e},$$

Krippendorfs Alpha. D_o is the observed disagreement. D_e, is the expected.

Reliability: Confusion matrix

167	0	0	0	15	0	0
5	104	0	0	12	0	5
0	0	30	4	1	0	0
0	7	4	118	0	0	0
18	9	4	0	506	1	4
0	1	1	0	0	28	0
4	3	0	0	6	1	127

Improving reliability and validity

- Clear and concise theoretical definition

- Efficient Coding Scheme
  - Definition
  - Paradigmatic Cases
  - Bordering Cases
  - Negative Cases

Example: Political Manifesto Project

https://manifesto-project.wzb.eu

The Manifesto Project provides the scientific community with parties' policy positions derived from a content analysis of parties' electoral manifestos. It covers over 1000 parties from 1945 until today in over 50 countries on five continents. The DFG-funded MARPOR project continues the work of the Manifesto Research Group (MRG) and the Comparative Manifestos Project (CMP). On this website you find the Manifesto Project Dataset containing the parties' policy preferences generated by the project. You also find coded and uncoded election manifestos of the parties in the dataset as well as information and links to many applications for the dataset, related projects and publications etc.

Example: Political Manifesto Project

https://manifesto-project.wzb.eu

Military: Positive

The importance of external security and defence. May include statements concerning:

- The need to maintain or increase military expenditure;
- The need to secure adequate manpower in the military;
- The need to modernise armed forces and improve military strength;
- The need for rearmament and self-defence;
- The need to keep military treaty obligations.

Computational Content Analysis

Discovery	Clustering - text similarity Topic Modelling Text summarization
Exploration	Word/sentence similarity search Computer-Assisted Keyword and Set Discovery
Classification	Natural Language Processing as Measurement Device Supervised Learning Rulebased classification

Category development: Model-in-the-loop

- Using machine learning to improve the process of **discovering** categories and the **development** of valid category schemes.

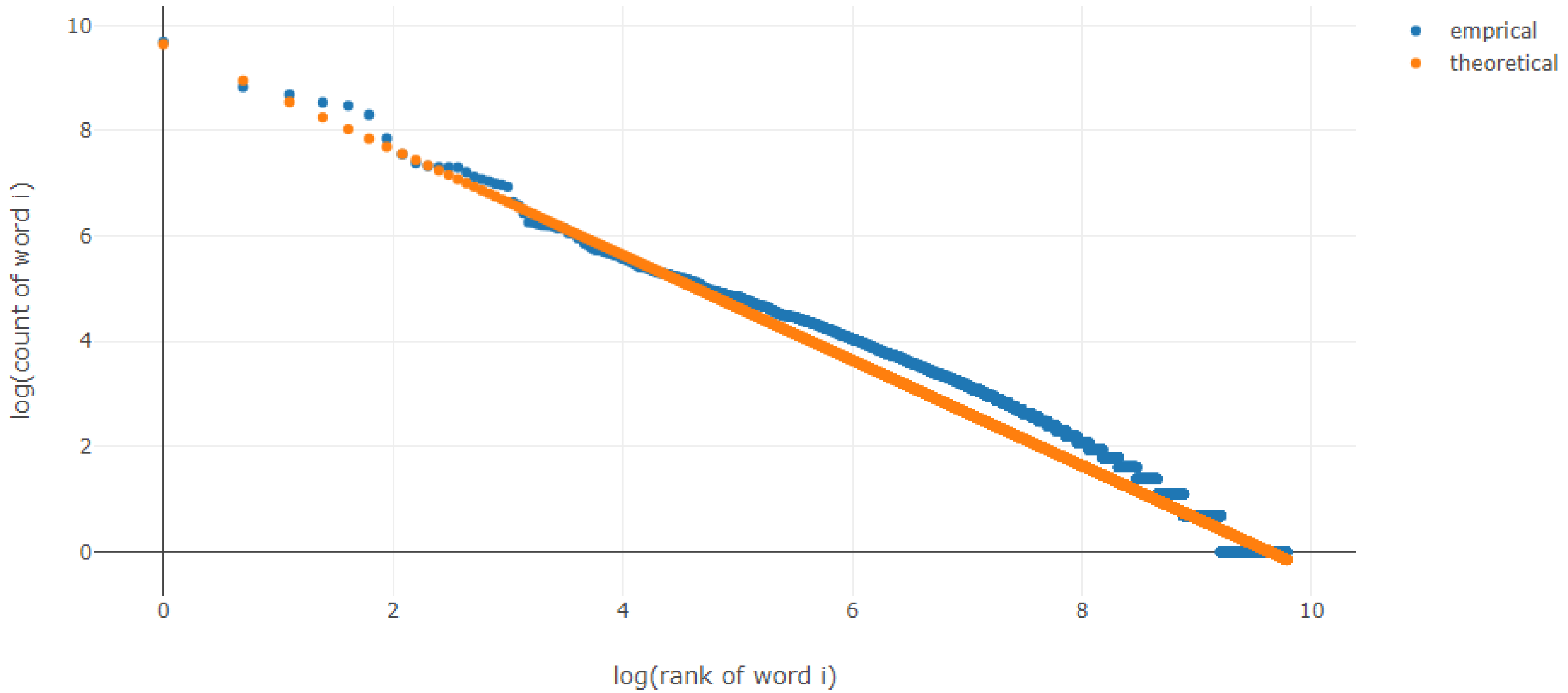
GOALS

- Discovery of interesting categories
- Understanding of corpus and its context
- Theoretical understanding of categories
- Good coding instructions:
  - Variety of manifestations,
  - Instructive cases: Exemplary cases, hard to spot / ambivalence, and bordering cases.

Learning about your data: Discovery

- Large Corpus of Text: What is in your data?
- Random sampling is not viable.
  - To much variations.
  - Everything is rare.

Zipf Law →



Learning about your data: Discovery

- Large Corpus of Text: What is in your data?
- Random sampling is not viable.
  - To much variations.
  - Everything is rare.
- Use Clustering and inspect representatives and summarizations.
  - Top documents.
  - Top words and phrases.

Learning about your data: Text Clustering

Agglomerative clustering: Agglomerative clustering: Hard-assigned and Hierachical (i.e. Multiple Solutions)

1. Transform documents to vector (tf-idf, bow, language model encoding).

Learning about your data: Text Clustering

Agglomerative clustering: Agglomerative clustering: Hard-assigned and Hierachical (i.e. Multiple Solutions)

1. Transform documents to vector (tf-idf, bow, language model encoding).
2. Compute similarity between (all) documents.

Learning about your data: Text Clustering

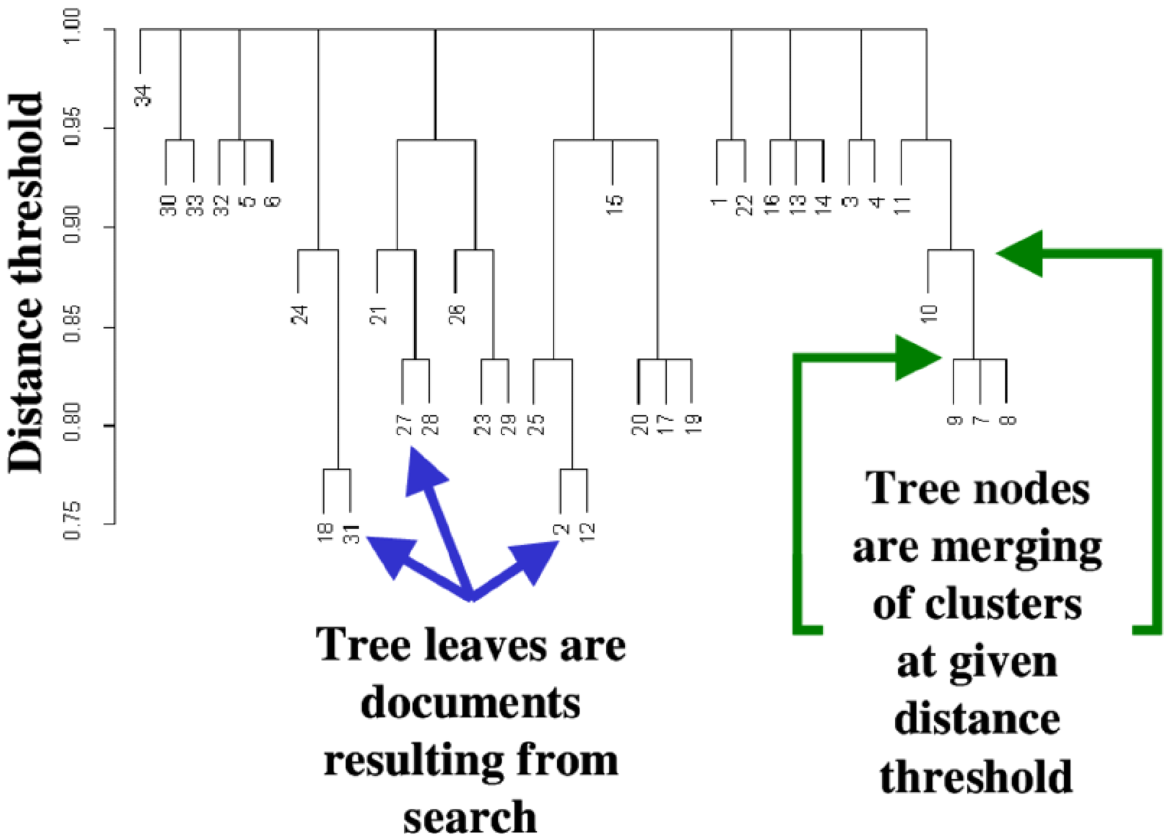
Agglomerative clustering: Hard-assigned and Hierachical (i.e. Multiple Solutions)

1. Transform documents to vector (tf-idf, bow, language model encoding).
2. Compute similarity between (all) documents.
3. Create dendogram by joining all pairs sequentially, ordered by their similarity score.

Learning about your data: Text Clustering

Agglomerative c -assigned and
Hierachical (i.e. I

1. Transform doc
2. Compute simi
3. Create dendo ered by their
similarity score.



Learning about your data: Text Clustering

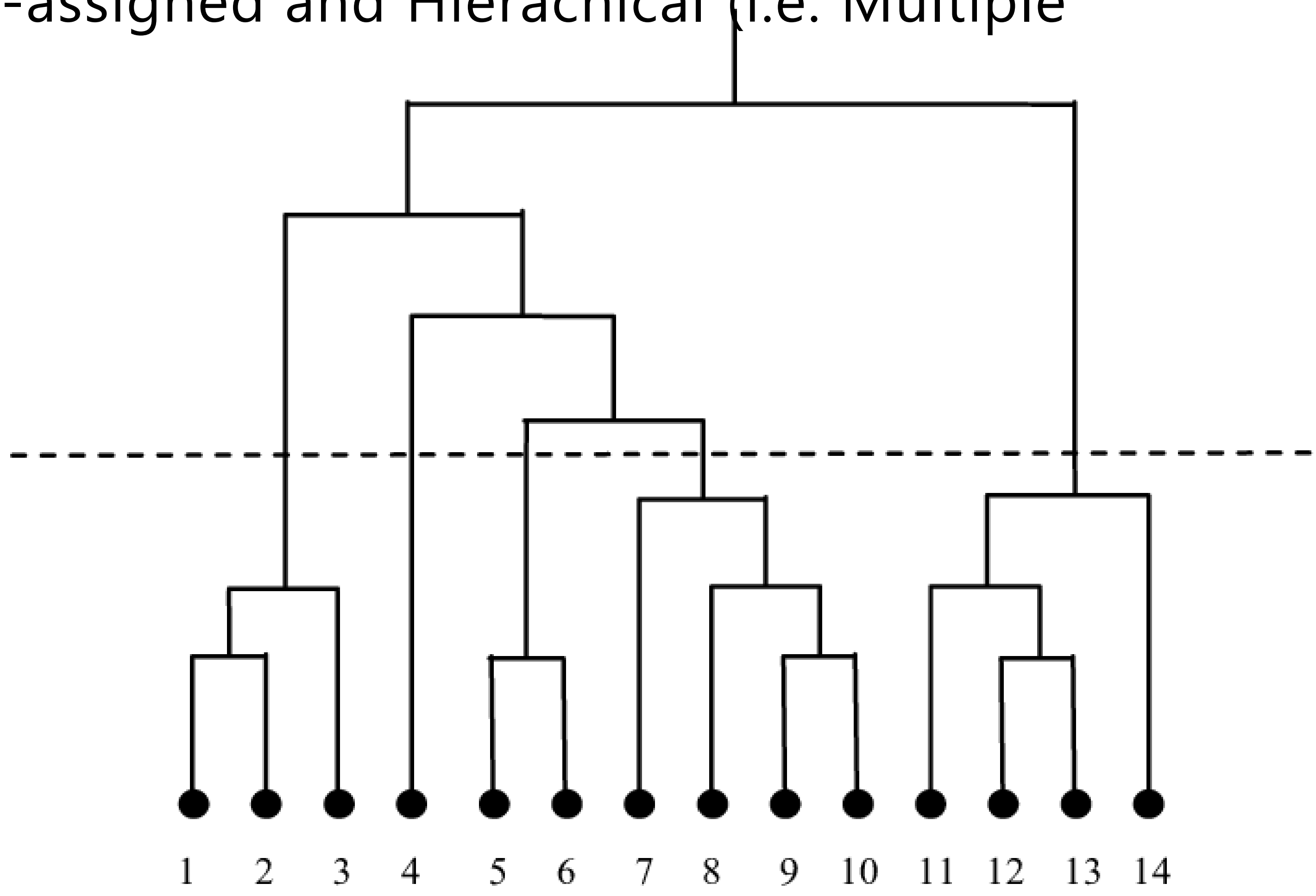
Agglomerative clustering: Agglomerative clustering: Hard-assigned and Hierachical (i.e. Multiple Solutions)

1. Transform documents to vector.
2. Compute similarity between (all) documents.
3. Create dendogram by joining all pairs sequentially, ordered by their similarity score.

4. Choose cut in the dendogram.

Learning about your data: Text Clustering

Agglomerative clustering: Hard-assigned and Hierachical (i.e. Multiple Solutions)



Learning about your data: Text Clustering

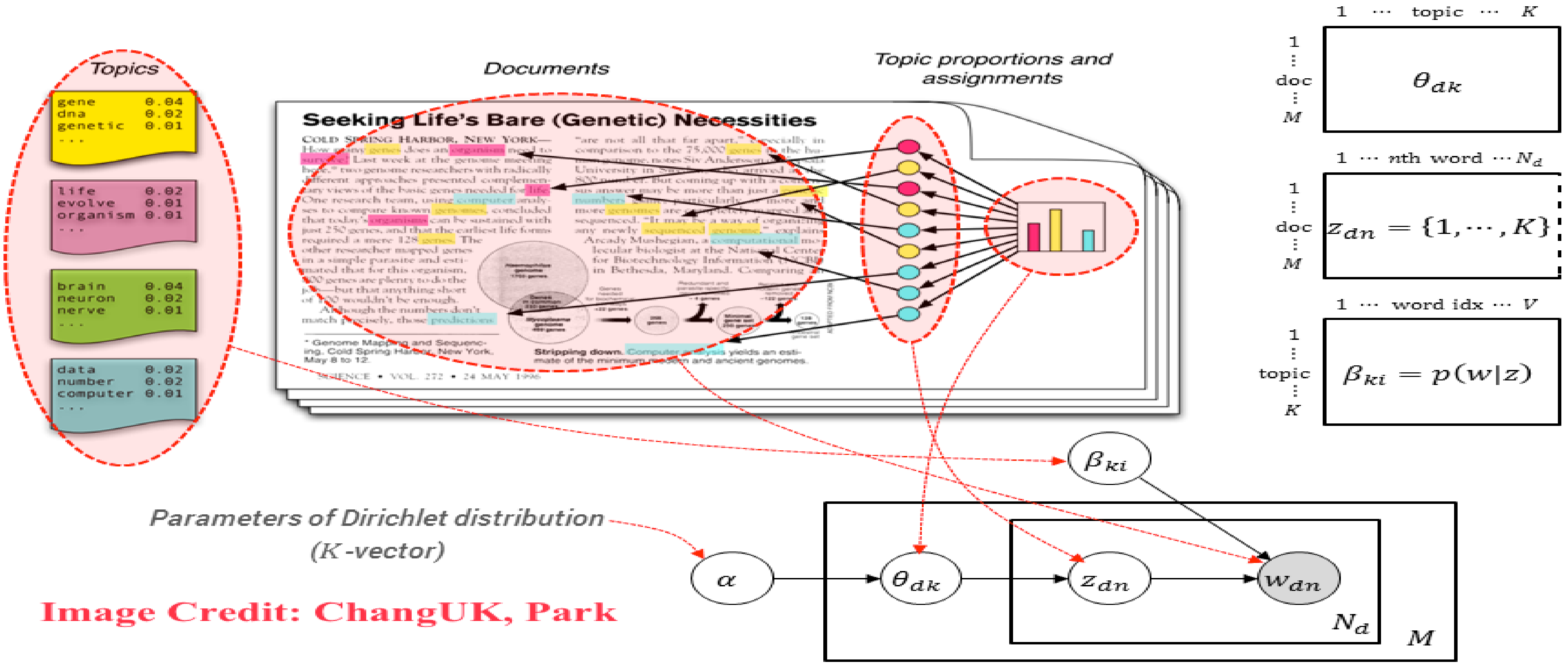


Image Credit: ChangUK, Park

Learning about your data: Text Clustering

TOPIC MODELLING

- Widely used tool in both research and industry.
  - powering search algorithms, document retrieval, and recommendation systems.
- Used for both measurement and discovery in the Social Sciences.

Pros

*** Mixed membership model: I.e. documents and words are softly assigned to multiple clusters.**

- Praised for its inductive and datadriven properties.
  - I.e. we did not come to the data with preconcieved theoretical ideas about what exists and what is important.
- Beyond atomized words, and can handle polysemi of words.
  - But still based on the BOW assumption.

Learning about your data: Text Clustering

Mixed-membership models: Topic Models.

- 1.generation: Latent Dirichlet Allocation, Structural Topic Model
  - Specifying Hyperparameters: No. of topics, α and β determining the degree to which words and documents are multi memberships.
  - Many assumptions about the generative process, which constrains the solution.
- 2. Hierachical Stochastic Block Model. (Gerlach et. al 2018)
  - Locates no. of topics automatically.
  - Much more flexible and sensitive especially to imbalance in cluster sizes.

Generative models (1)

- Define a model that you believe describe the data generation process.
  - E.g. which parameters determine the probability of a network tie,
  - Word in a document.
- Define the variables and their dependencies.
  - Network: Same school, ethnicity, culture, gender.
  - Words: Mood, speaker, social situation.

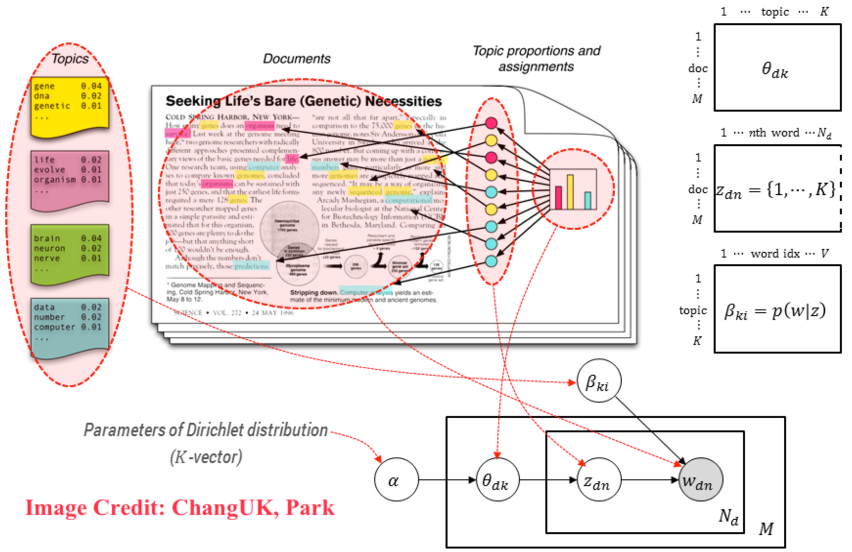


Image Credit: ChangUK, Park

Generative models (2)

Naive Bayes : p(x,y)

Simple generative model for the probability of a drawing "word" given a categorical variable.

** Example **

p(w=Yes | y=tired)

Following me every morning we can observe and approximate p(yes|tired)

- Using bayes rule
- p(y|x) = p(x|y)*p(x) / p(y)
- We observe p(x|y), p(y) and p(x).

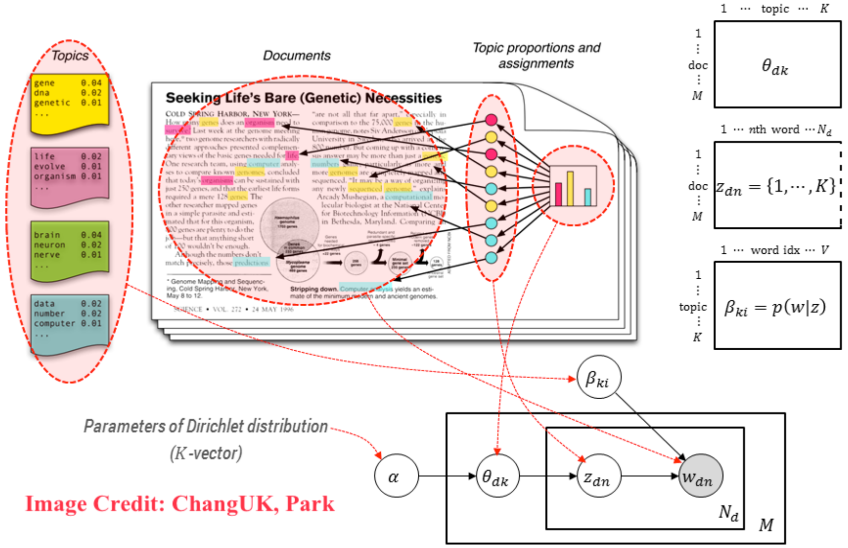


Image Credit: ChangUK, Park

Generative models (3)

LDA

- Variables in the generative model does not have to be observable.
  - They can be **latent** (similar to neural networks)

Model definition:

- The probability of a "drawing" a word is dependent on the topic.
  - Topics are **Latent** - i.e. not observed.
  - Topics are distributions of word probabilities.
    - Words can be present in more than one topic.
- Documents consist of multiple topics.

** Can be extended with any latent structure as well as known variable **

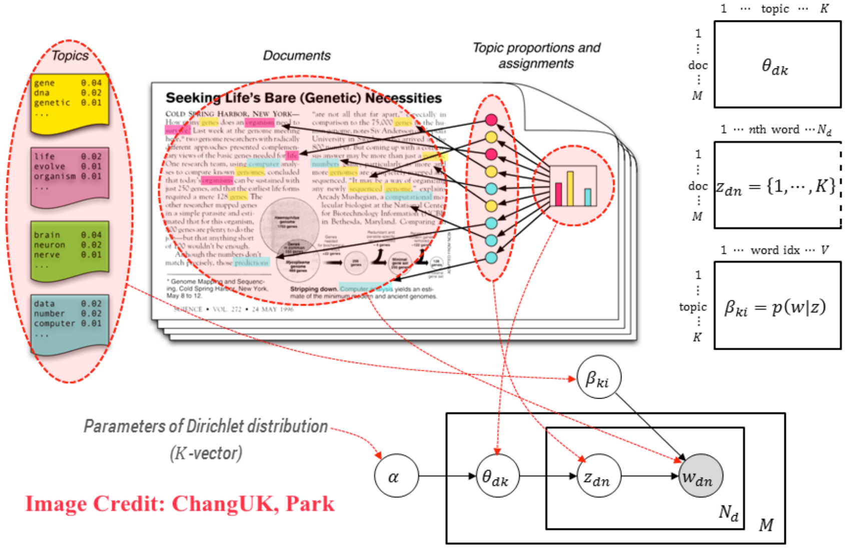


Image Credit: ChangUK, Park

Models with unrealistic assumptions can produce wildly misleading results

Lanchichetti et. al. 2014:

LDA collapses different languages into the same cluster given a wrong prior α prior.

Complex models are hard to fit

- Instability of solutions and local minima.
  - (Lancietti et al. 2014; Chuang et al. 2015, Roberts et al. 2016, Wilkerson and Casas 2017; Gentzkow et al. 2017: 27; Agrawal et al. 2018),

Learning about your data: Discovery

Important capabilities

Variety

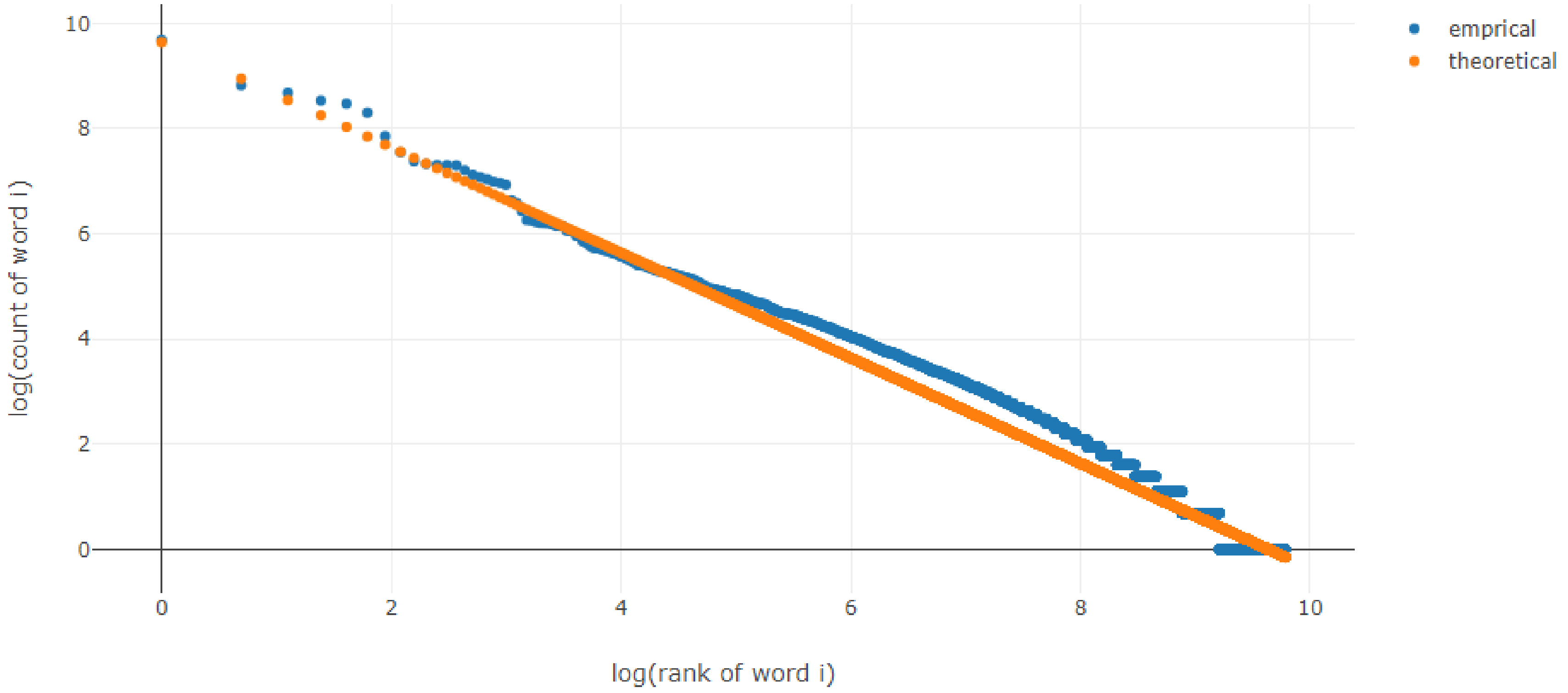
Does it cover style, topicality, issues, stances, situations?

Learning about your data: Discovery

Important capabilities

Scale

Can it cover small and big topics (Classic LDA style topic modelling fails here)



Learning about your data: Discovery

Important capabilities

Variety

Does it cover style, topicality, issues, stances, situations?

Scale

Can it cover small and big topics (Classic LDA style topic modelling fails here)

Language and Context understanding

Word usage, Negations, grammar, compositionality, sequentiality, context, implications,

Learning about your data: Discovery

Capabilities

Input X Algorithm

- Bow or Language model encoding
- Agglomerative or Probabilistic

Summarization

Inspect top documents and top words.

-- initial category definition.

Learning about your data: Exploration

- *Define-inspect-refine*
  - discovered (top) words → Document set → add words.

  - Example.
  - Broad at first to get at the variation. Produce Subcategories later (by subclustering).
- Query building: Statistical similarity search: E.g. Word2Vec.
- Document based neighbor search: Document
  - Similarity / distance measures:
    - dot product between two vectors.

- Model-in-the-loop queries:
  - King et al. 2017: "Computer-Assisted Keyword and Document Set Discovery from Unstructured Text"

Learning about your data: Exploration

- Model-in-the-loop: "Computer-Assisted Keyword and Document Set Discovery from Unstructured Text"

- Start with an initial query (set of keywords).
- Classify documents using these.
- Train a Classifier to match the query result.
- Inspect most predictive features (e.g. largest coefficients in the logistic regression model).
- Repeat

Word Embedding Based Similarity Search

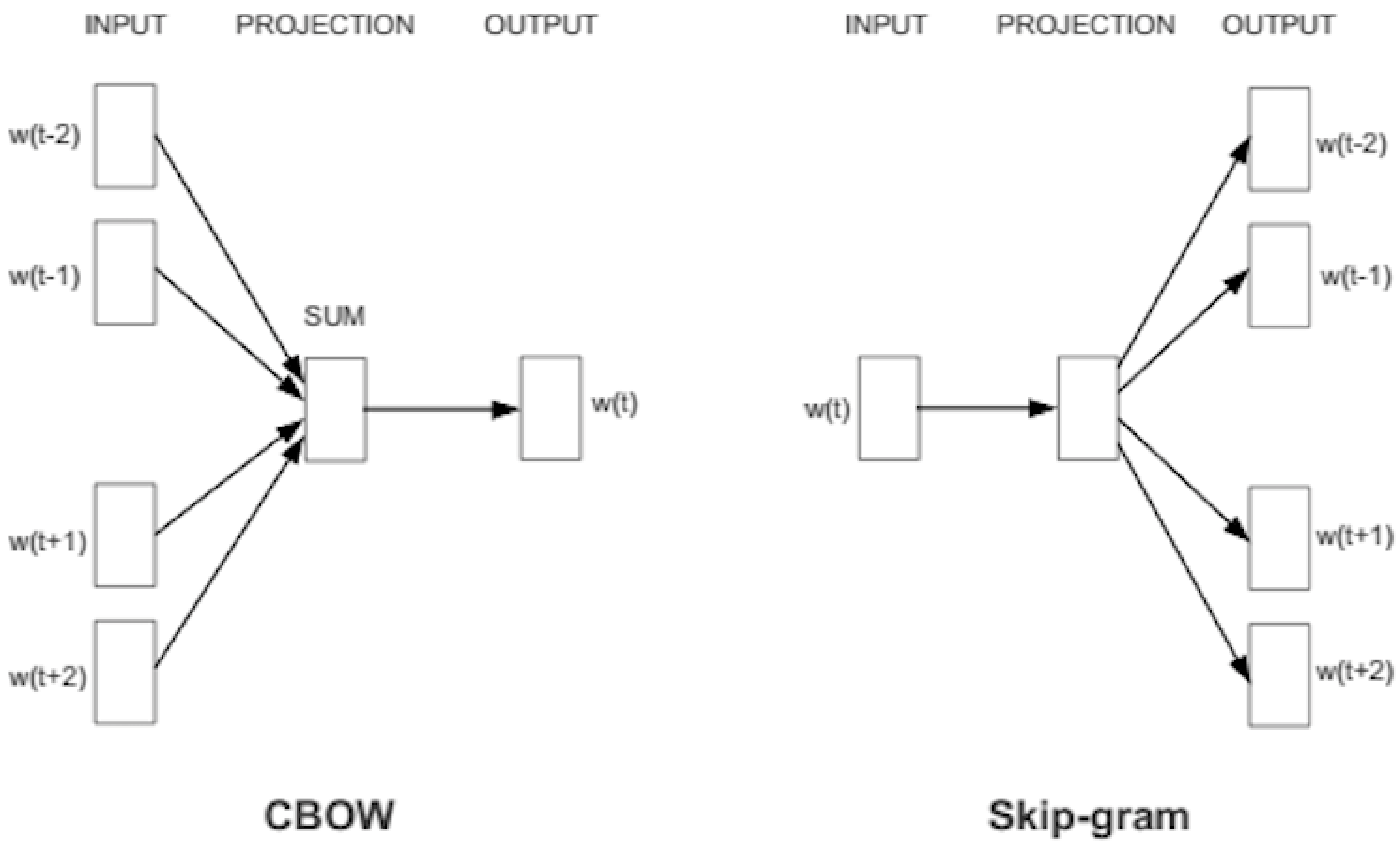
Word2Vec

- Mikolov et al. 2013: "Efficient Estimation of Word Representations in Vector Space"

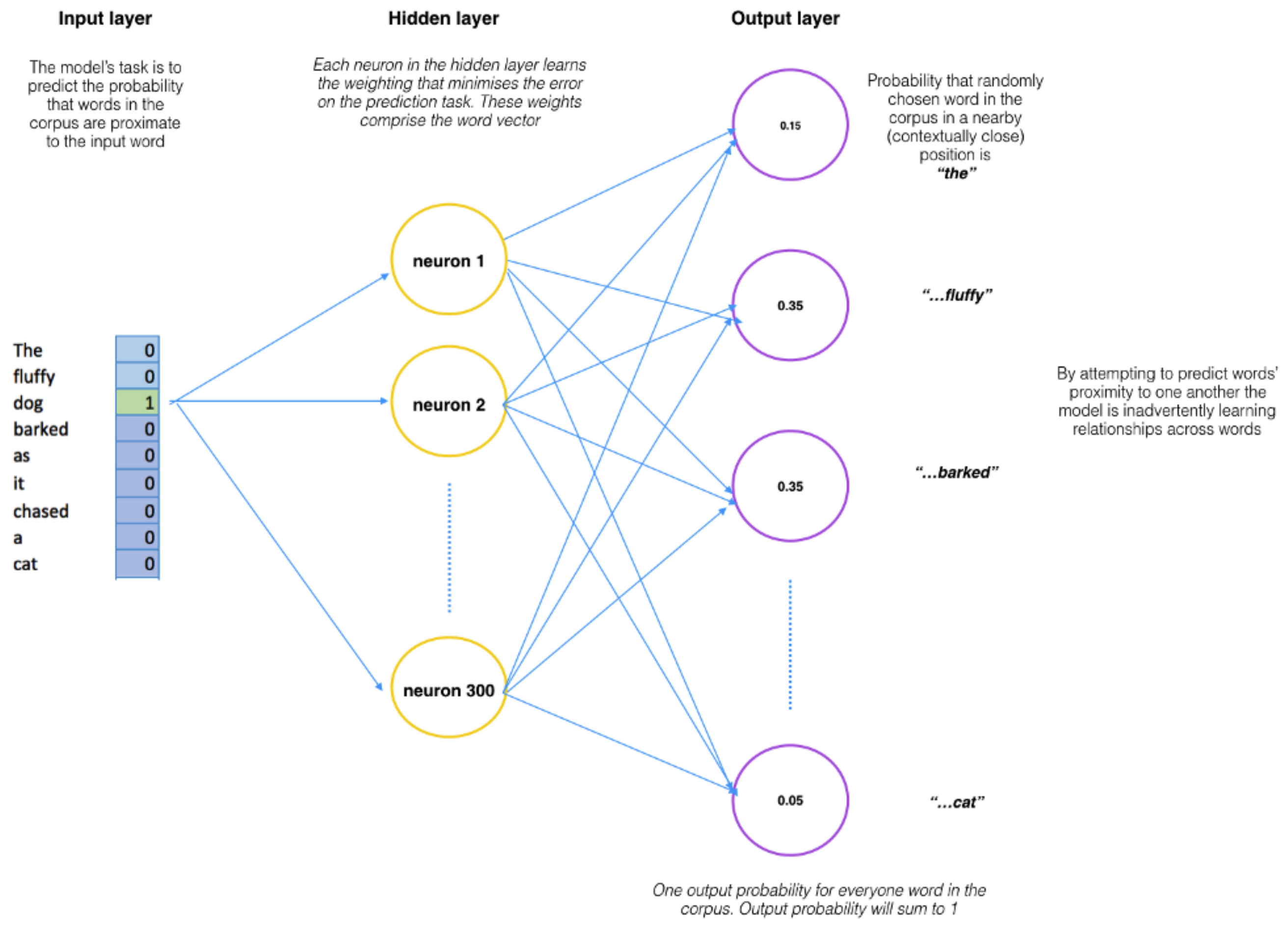
Self-supervisory tasks : Original Transfer Learning / Language Model

- Skip-Gram: -Given a word: predict the previous and the next *k* words.
  - Embed words in a contionous vector space and maximize cross-product between "neighbor" words.
    - Essentially an efficient way of processing the NxN co-occurrence matrix.
- CBOW: Contionous Bag of Words.
  - Given a set of context words (before and after): predict the middle word.
  - Maximize the cross-product between the middle word and the sum of its neighbors.

Word Embedding Based Similarity Search

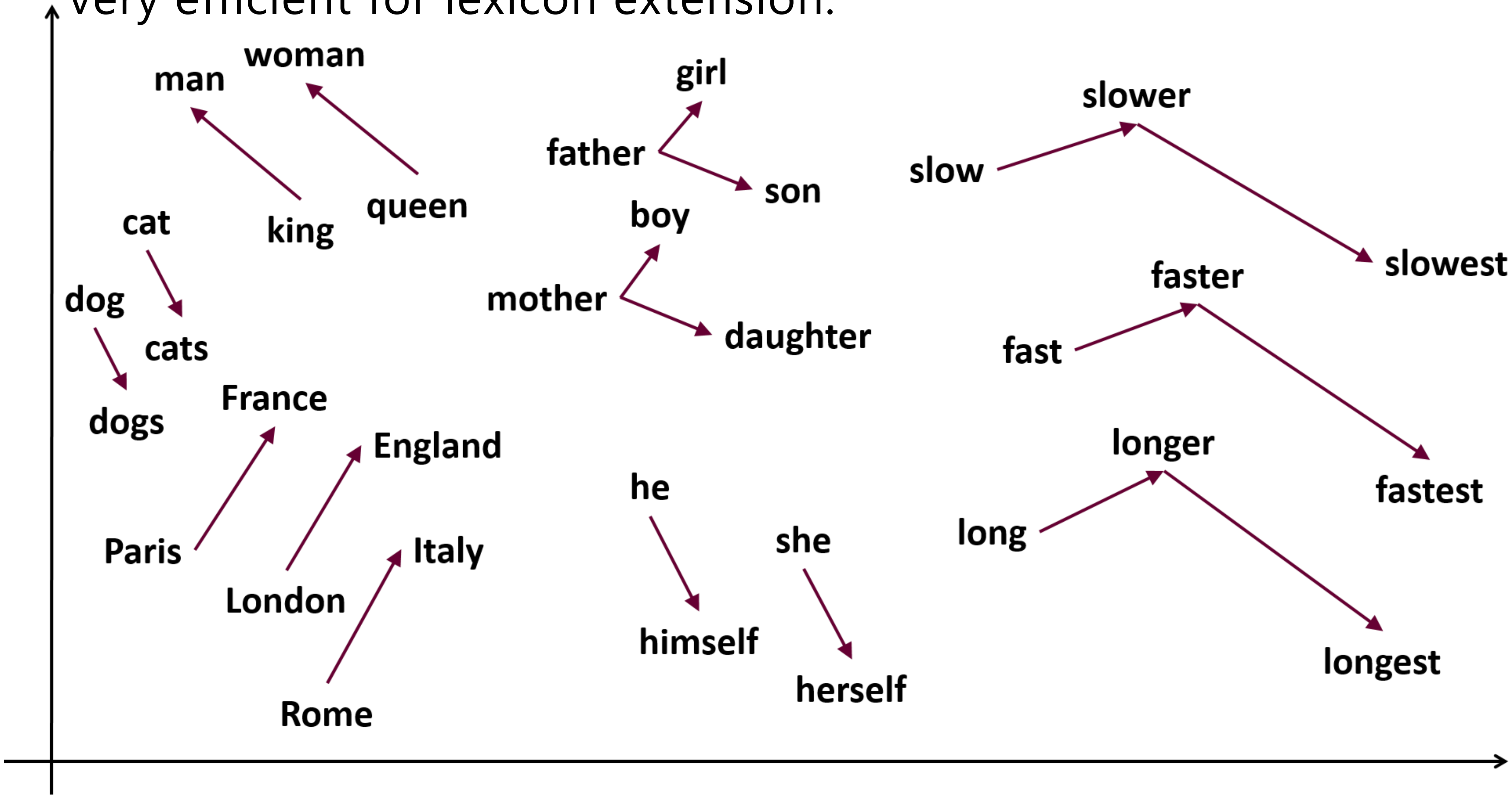


Word Embedding Based Similarity Search



Word Embedding Based Similarity Search

- Learns analogies,Semantic syntactic information, Topical information → very efficient for lexicon extension.



Word Embedding Based Similarity Search

- Does not handle polysemy and multiword unless preprocessing is done right
  - each token has a single representation. "$ bank" and "river bank" share one representation.

Word Embedding Based Similarity Search

Levy, Goldberg and Dagan 2016: "Improving Distributional Similarity with Lessons Learned from Word Embeddings“

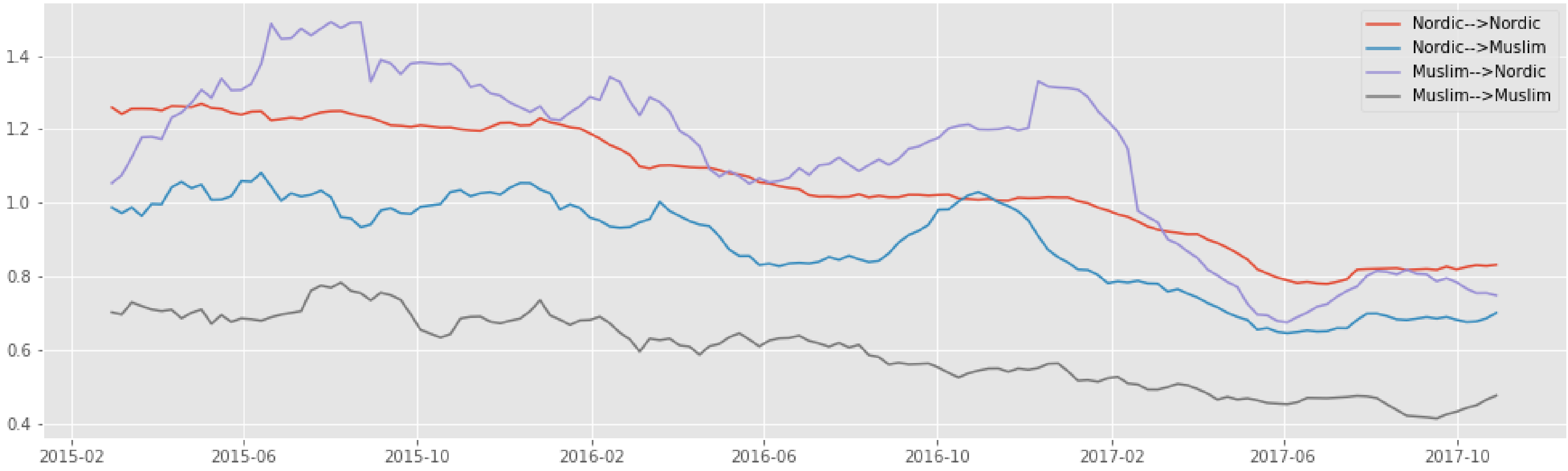
- - More classic methods of dimensionality reduction (e.g. SVD) are comparable, once Word2Vec "tricks" are applied.
  - - dynamic context window instead of co-occurrence based on document.
- - smoothing methods (counts are raised to the power of 0.75)
- - sampling methods (intelligent removal of very frequent words)
- - add context vector (represent each word as its own vector plus average context vector)

But new "Online" Methods (stochastic gradient descent) are far more efficient for estimation in Big Data settings.

Refining Categories: Specifying variation

- Broad set of Keyword heuristics.
- Broad Document Set.
- Subcluster documents.
- Inspect subclusters to locate Paradigmatic Cases, and Bordering Cases.
- Create descriptions.

Cross-ethnic hostility in Denmark



First outcome
(1) Is this a discussion / polemic between strangers
(2) Friendly or tag interaction,
(0) or unknown / seems irrelevant?

<u>If: (1) Is this a discussion / polemic between strangers</u>

Does the comment seem to express:
(1) Agreement.
(2) Disagreement,
(0) Unknown

<u>If: (2) Friendly or tag interaction eller (0) or unknown / seems irrelevant?</u>

- Write a note

<u>If 1 or 2 or 0:</u>
Evaluate the Relational Content / Emotional Tone - towards the other:
(1) Friendly, Polite, Positive (this includes a polite and respectful disagreement).
(2) Hostile - Negative / Conflict,
(0) Not Negative - i.e. Neutral?

<u>If (1) or (0): Friendly, Polite, Positive</u>

Evaluate relational content of the comment:
(1) Expressing full support
(2) Expressing some / conditional support.
(3) Polite expression of disagreement - i.e. seeking compromise, acknowledging the others argument, reasoning tone.
(4) An attempt at persuasion. - As opposed to just expressing an opinion or taking a stand in a conflict
(0) Something else.
If more then one separate by ,

<u>If (2): Hostile - Negative/Conflict:</u>

Evaluate relational content of the comment:
(1) Aggressive
(2) Passive-aggresive
(3) Condenscending
(4) Sarcastic
(5) Trolling
(6) Personal attack (i.e. ad hominem attack, questioning authority)
(7) Provocative.
(0) Something else

<u>And hereafter:</u>

Evaluate the Explicit(!) Subjective Sentiment expressed- Not Directed(!) towards the other -:
(1) Anger
(2) Sadness
(3) Despair.
(4) Joy, Happiness
(0) No subjective emotional content - Neutral (if the content is only relational).
(-1) Unknown.

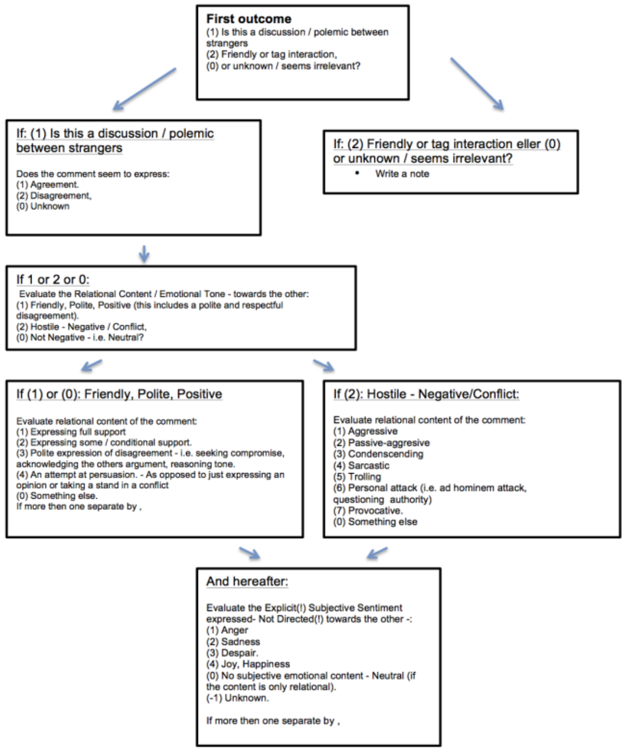
If more then one separate by ,

- Agreement

"Completely right. They are just waiting to become enough people. It is heard from more imams"

"God it is so stupid and greedy."

This reply expresses the same opinion as the comment/post that it is a reaction to. It is from this reasoning that it is coded as 'expressing full support'.

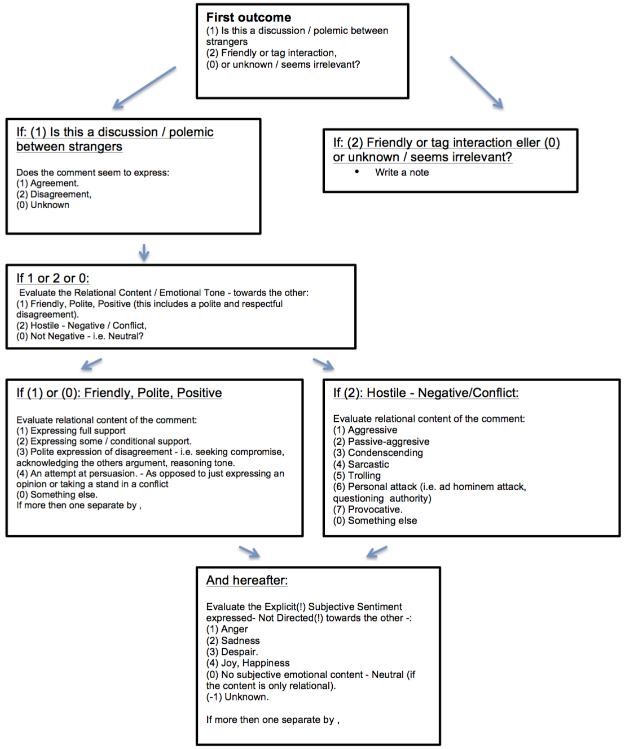


Condenscending

"Are you Danish? I'm doubting it from reading what you write?"

"(name) so that was today's most far out reply. It didn't make any sense at all. I'm thinking redwine?"

"But then someone believed him, so your statement has been undermined. have a really nice day"



THANK YOU FOR LISTENING