NIHR | Maudsley Biomedical Research Centre

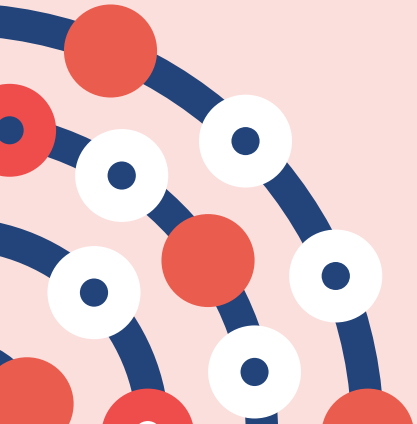
Modelling language: documents

Angus Roberts, Senior Lecturer in Health Informatics
Institute of Psychiatry, Psychology and Neuroscience
King's College London

Representing language

- How can we represent meaning in language?
- Symbolic, rationalist approaches
- Empirical approaches to representing documents
  - Bag-of-words
  - TF-IDF

Symbolic, rationalist approaches



Symbolic NLP

- If we want to manipulate language computationally, we need to represent it in some way
- In rule-based, symbolic NLP, we can consider language to be represented as strings of characters
  - A string of characters has little inherent meaning
  - We match these strings with expressions (rules, grammars) that define some pattern of characters
  - These rules give meaning to our strings
  - The approach can be extended to POS and other features
- The thinking is that language can be reasoned about in some logical way, and that the structures of language could be can in to sets of rules
- Prolog, a logic programming language, is a typical approach

Prolog grammars

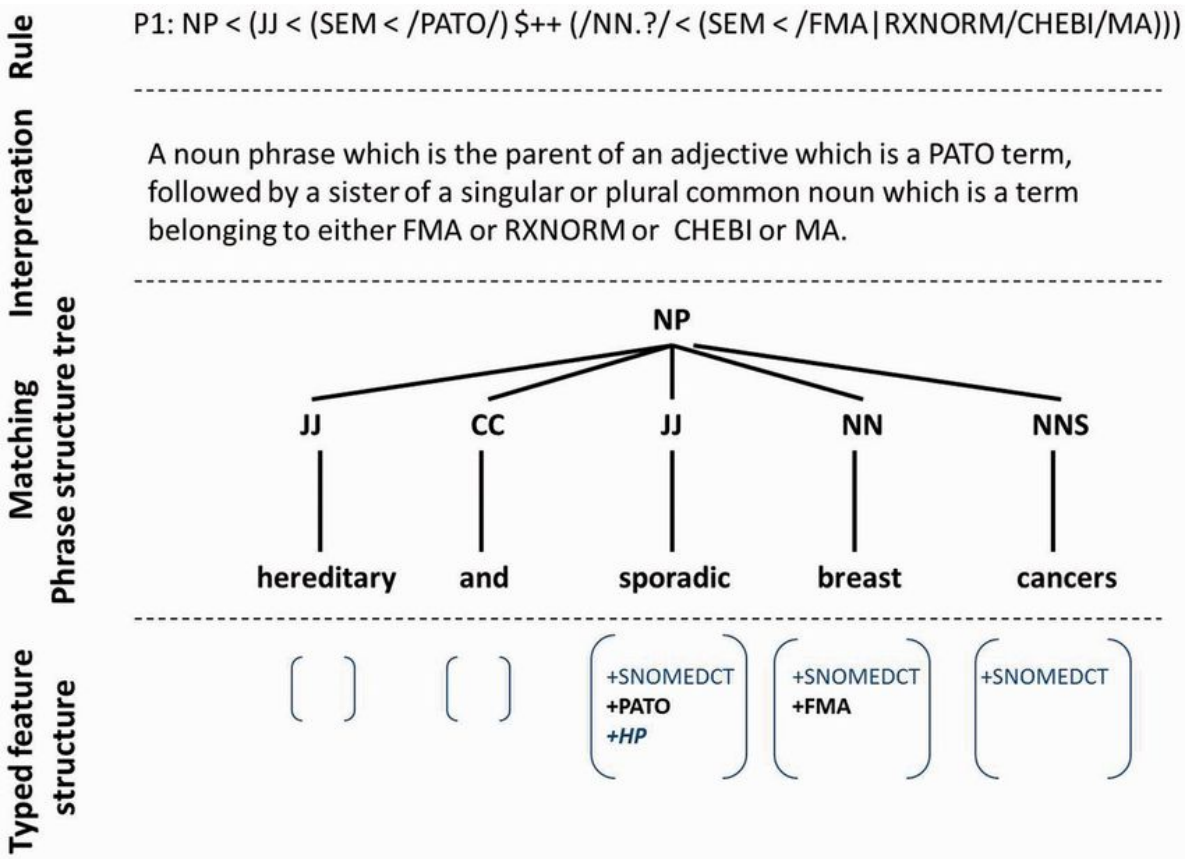
```
/* DOG_GRAMMAR.PL */

s --> np, vp.

np --> n.          np --> adj, n.     np --> adj, adj, n.
np --> det, n.     np --> det, adj, n.     np --> det, adj, adj, n.

vp --> v, np.     vp --> v, pp.

pp --> p, np.

det --> [the].     det --> [a].     det --> [an].

n --> [dogs].     n --> [fox].     n --> [jumps].

adj --> [quick].     adj --> [brown].     adj --> [lazy].

v --> [jumps].     v --> [runs].

p --> [over].     p --> [onto].
p --> [in].          p --> [under].
```

Credit: John Coleman, http://www.phon.ox.ac.uk/coleman

Grammars

- How many rules does it take to represent “grammatical” English?
- Do we have the same problem for all languages?
- What about medical language?
- Can we write grammars that capture not the syntax, but the semantics of language? i.e. the real-world categories that words relate to, and the relationships between them?
  - e.g. diseases, medications, anatomy?

Semantic grammars



Collier et al (2015). PhenoMiner: From text to a database of phenotypes associated with OMIM diseases. Database. 2015. bav104. 10.1093/database/bav104.

Empiricism

- Empiricism is at the heart of current NLP
  - Statistical models of language
  - How often do words appear?
  - In what context do they appear?
  - How many of a particular word appear in a piece of text?
- Is it disturbing that counting words in a document outperforms reasoning about the language?

Rationalism vs Empiricism

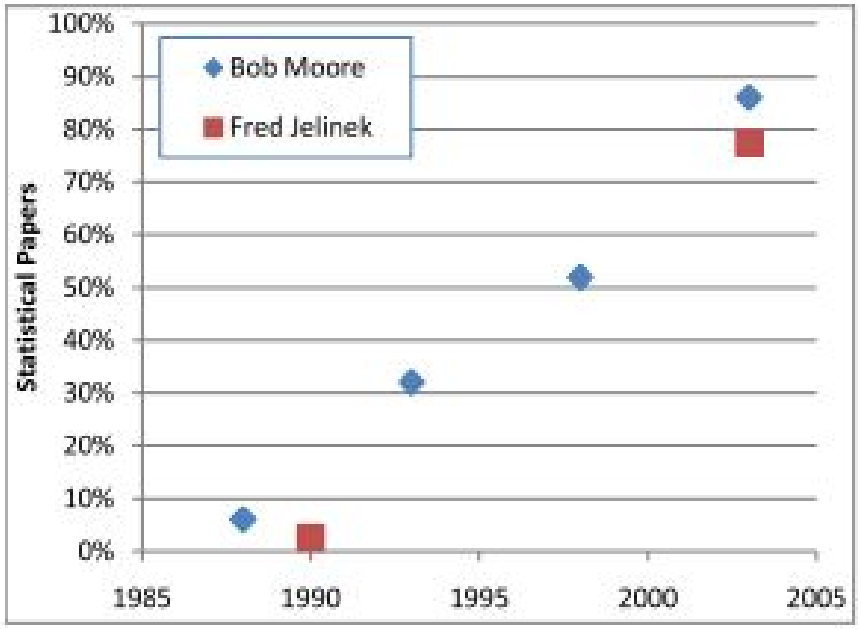
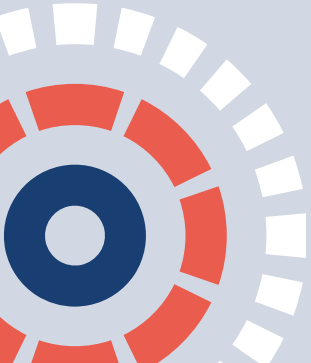


FIGURE 1 The shift from Rationalism to Empiricism is striking (and no longer controversial). This plot is based on two independent surveys of ACL meetings by Bob Moore and Fred Jelinek (personal communication).

- Church, LiLT Volume 2, Issue 4 May 2007

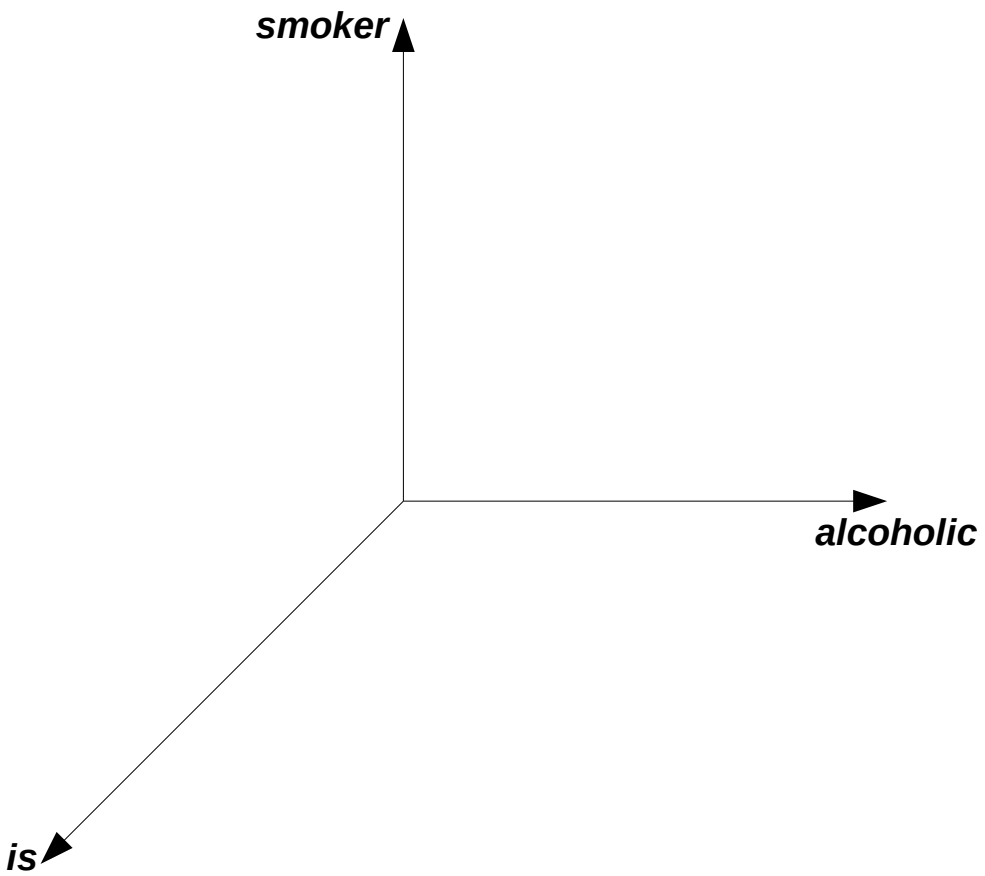
Empirical representations: bag-of-words



Bag of words

- Let’s consider four sentences:
  - he is a smoker
  - she is alcoholic
  - he is anxious
  - he is diabetic and diet controlled
- We will plot these along 3 dimensions:
  - smoker
  - alcoholic
  - is

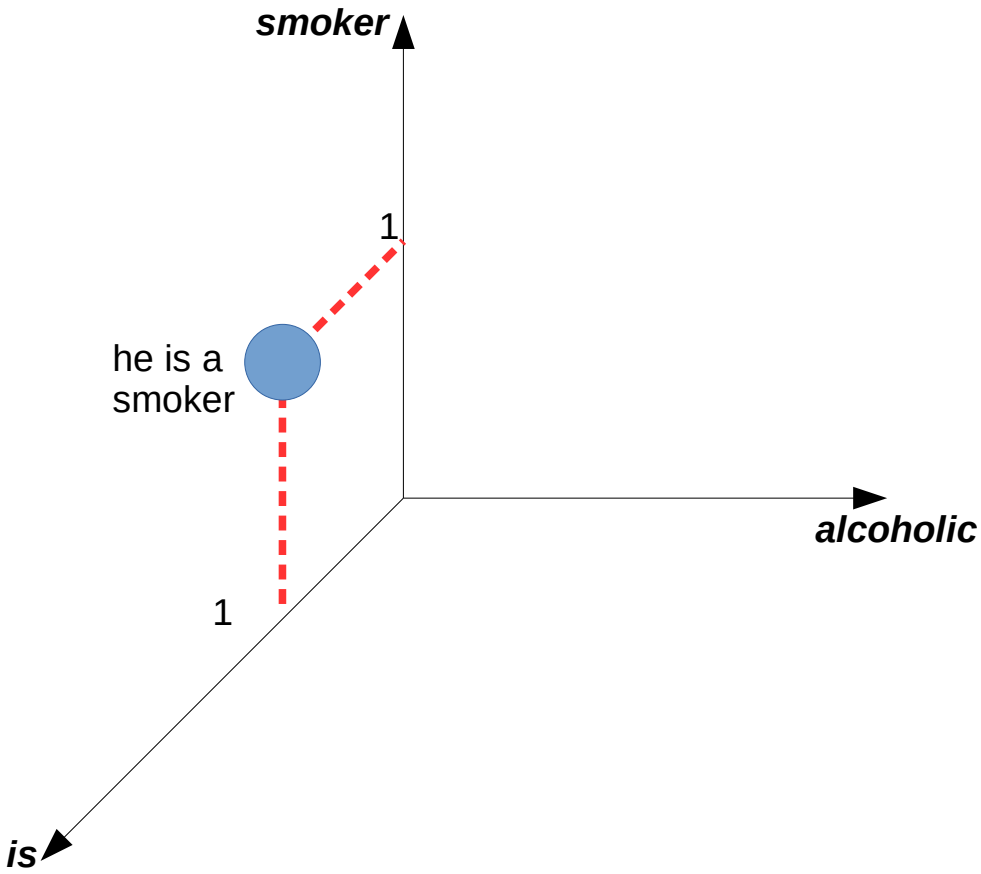
Bag of words



(After: Feature Engineering for Machine Learning by Amanda Casari, Alice Zheng. O'Reilley)

Bag of words

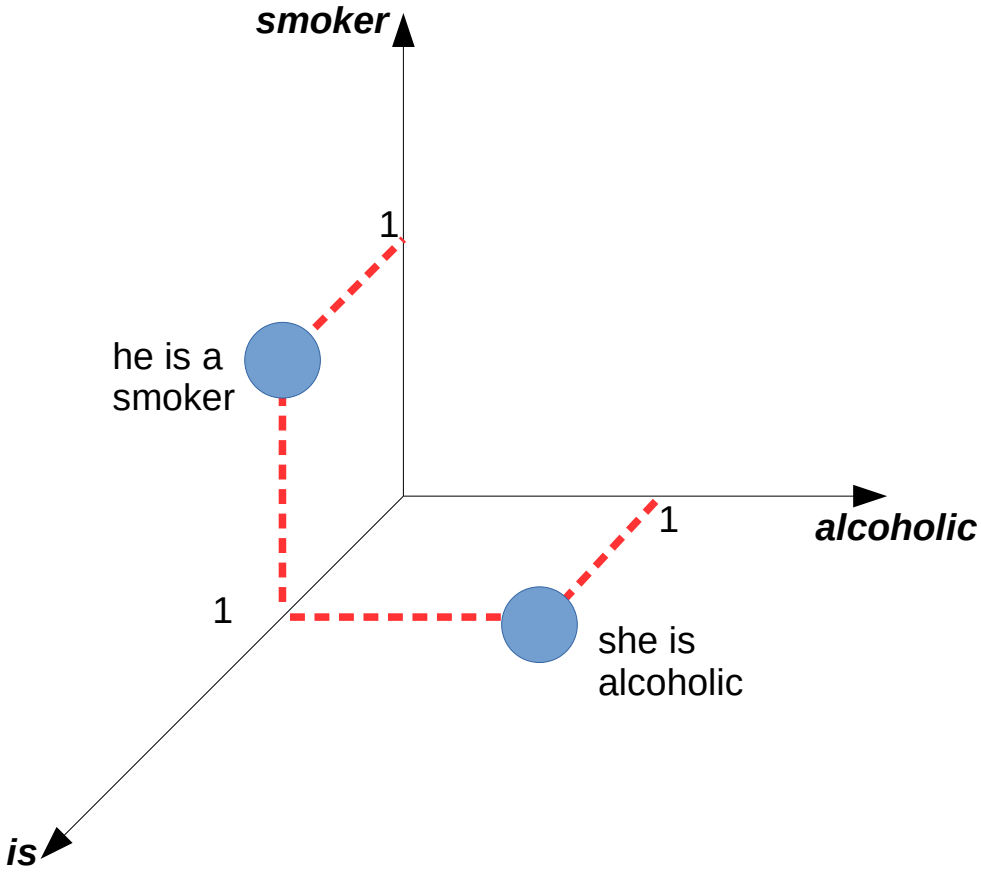
- he is a smoker



(After: Feature Engineering for Machine Learning by Amanda Casari, Alice Zheng. O'Reilley)

Bag of words

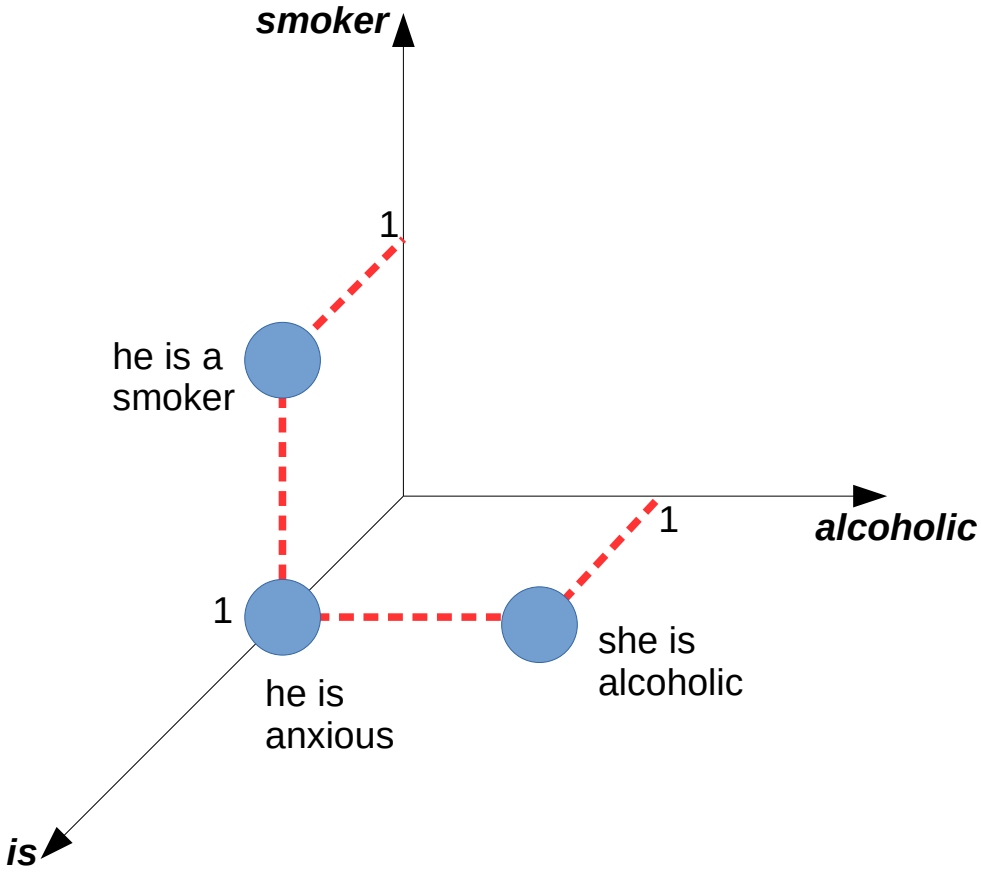
- he is a smoker
- she is alcoholic



(After: Feature Engineering for Machine Learning by Amanda Casari, Alice Zheng. O'Reilley)

Bag of words

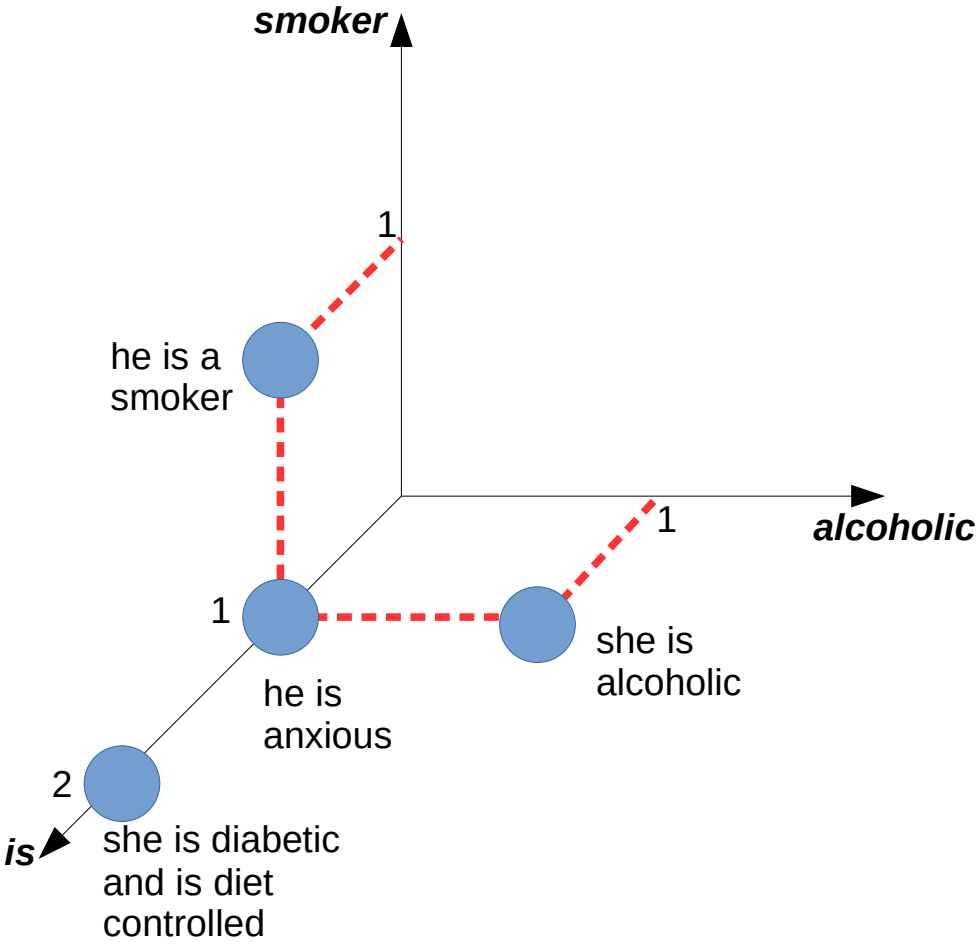
- he is a smoker
- she is alcoholic
- he is anxious



(After: Feature Engineering for Machine Learning by Amanda Casari, Alice Zheng. O'Reilley)

Bag of words

- he is a smoker
- she is alcoholic
- he is anxious
- she is diabetic and is diet controlled



(After: Feature Engineering for Machine Learning by Amanda Casari, Alice Zheng. O'Reilley)

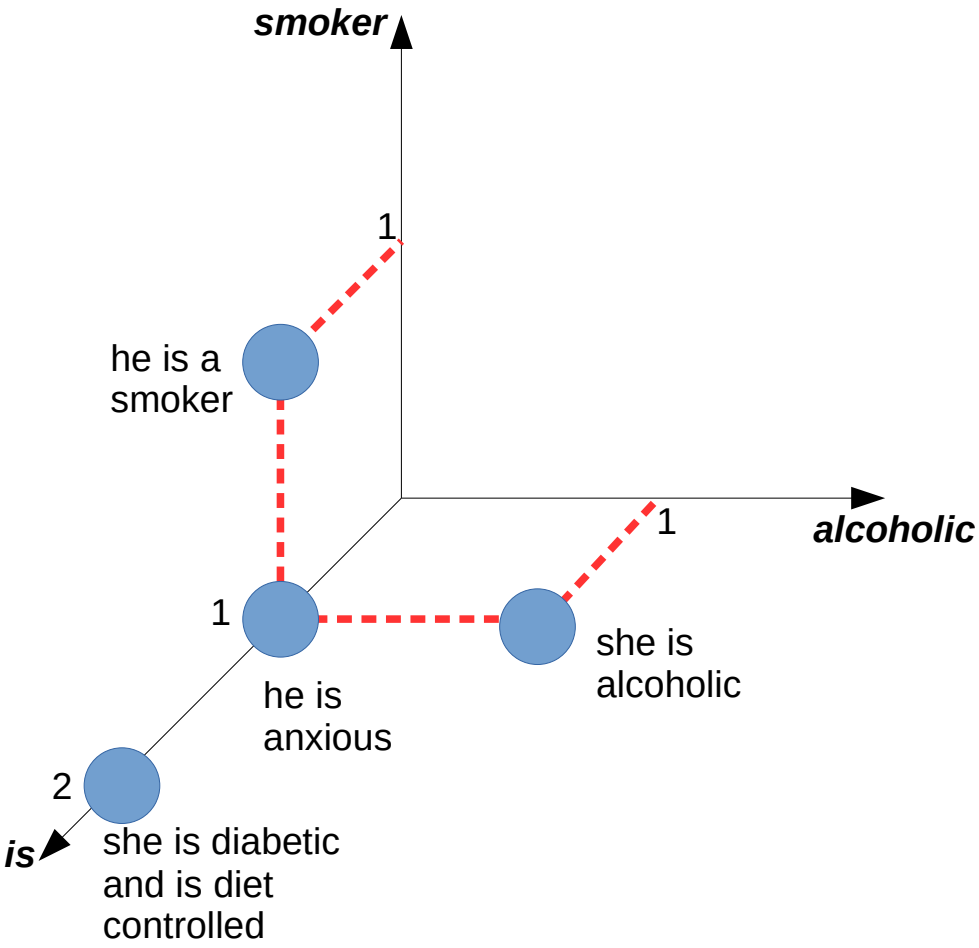
Bag of words

- Works surprisingly well on some problems
- But…
  - No word order: loss of context
    - “is this cancer” vs “this is cancer”
  - The Curse of Dimensionality: the power of our classifier reduces as the number of dimensions increases
  - Over fitting: given a low number of training instances relative to number of features
  - Important but less frequent words can have less of an influence than less important but more frequent words

Term frequency and inverse document frequency

Improving bag of words

- The occurrence of a rare word like “smoker” or “alcoholic” has as much influence on the vector as the occurrence of a common word like “is”
- Lots of occurrences of a common word (such as two mentions of “is” in one sentence) has a bigger effect than a more discriminating rare word.
- Note that we are just using “is” as an example, Often, we would deal with such highly frequent, so-called stop words, by filtering them out. TFIDF is still relevant for other frequent words.



TFIDF

- We can scale our BoW ***term frequencies***, multiplying each by a factor that accounts for how rare the term is – an ***inverse document frequency (idf)***

$$idf\ for\ word = \frac{number\ of\ documents}{number\ of\ documents\ containing\ word}$$

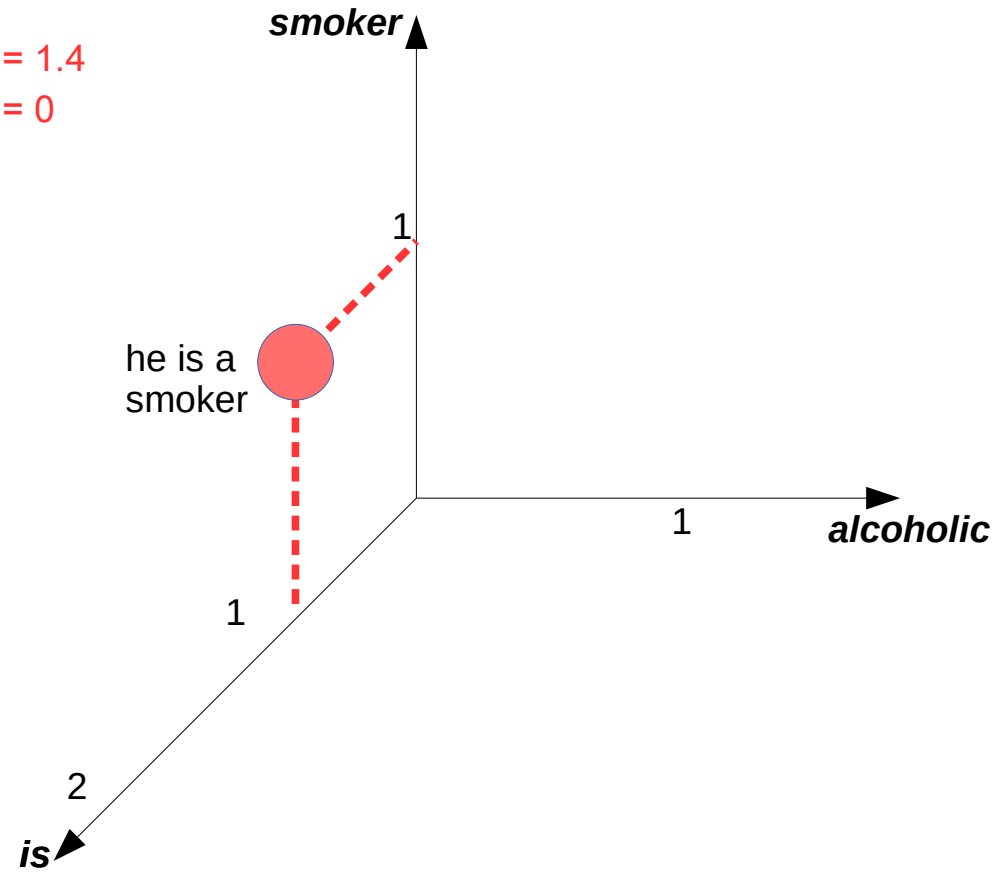
- Usually this is scaled further by taking the log
- The rarer a word, the higher idf
- The more common a word, the lower idf
- tf x idf scales the influence of each term accordingly

Improving bag of words

Document 1:

tfidf(smoker) = 1 * ln(4/1) = 1.4

tfidf(is) = 1 * ln(4/4) = 0

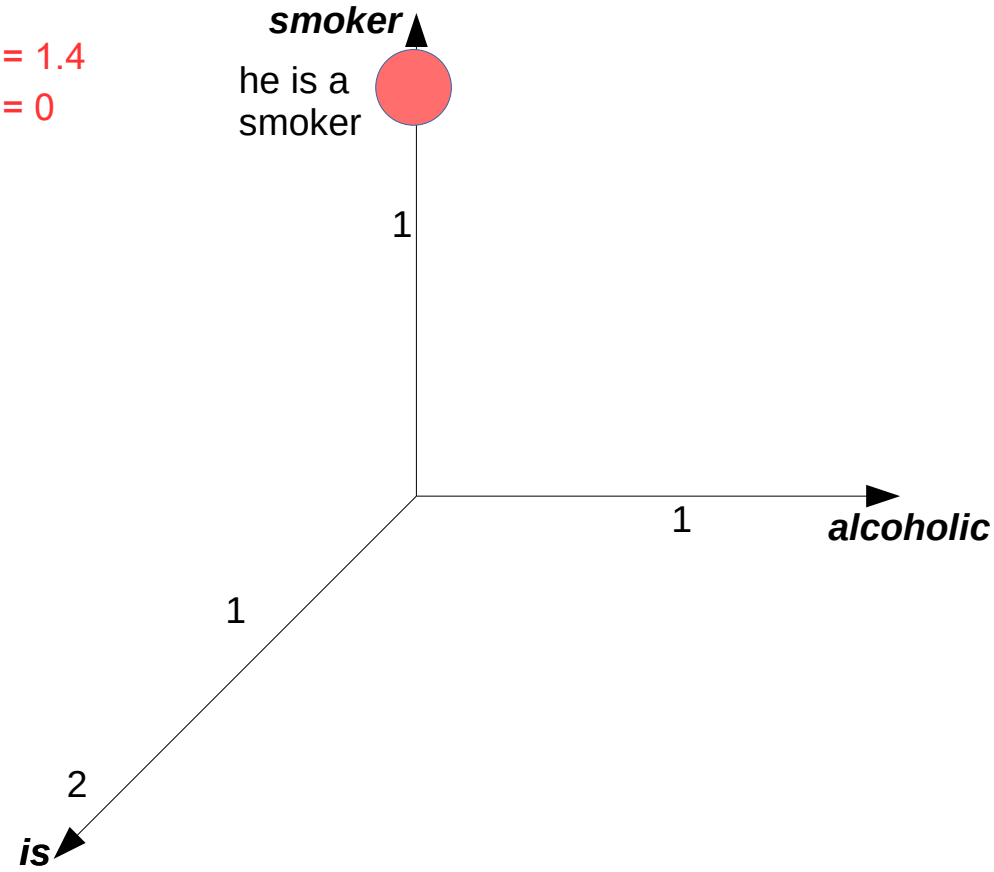


Improving bag of words

Document 1:

tfidf(smoker) = 1 * ln(4/1) = 1.4

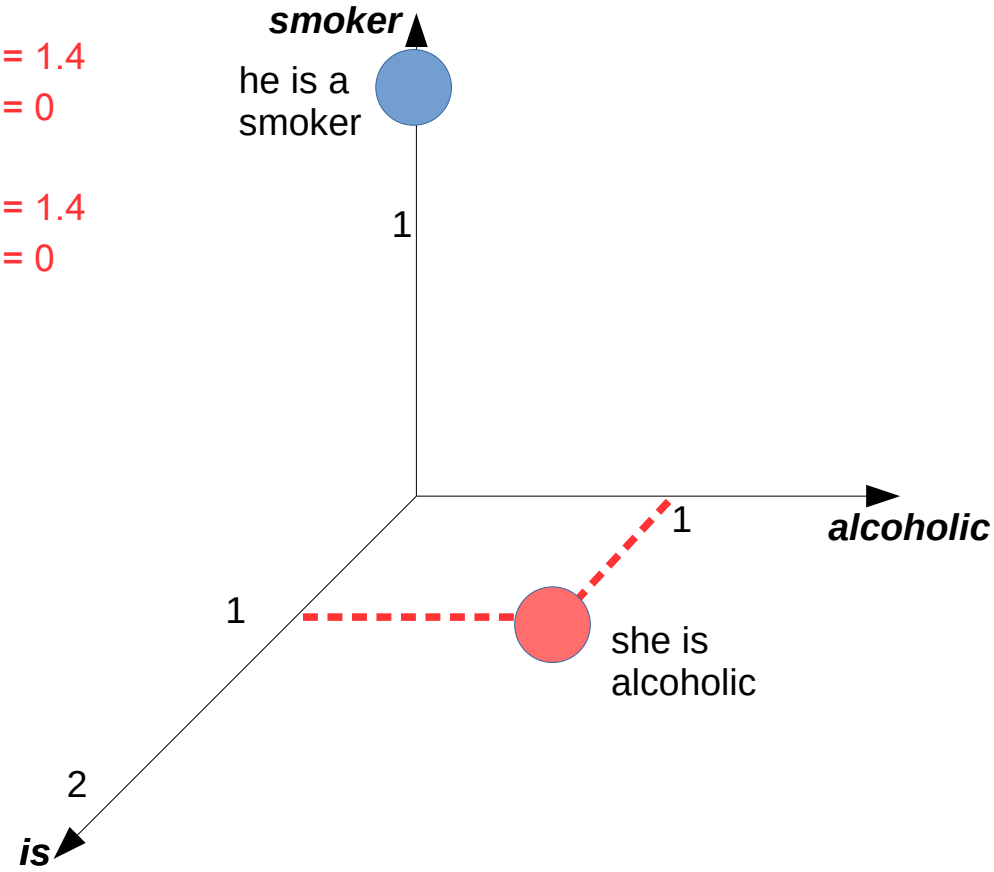
tfidf(is) = 1 * ln(4/4) = 0



Improving bag of words

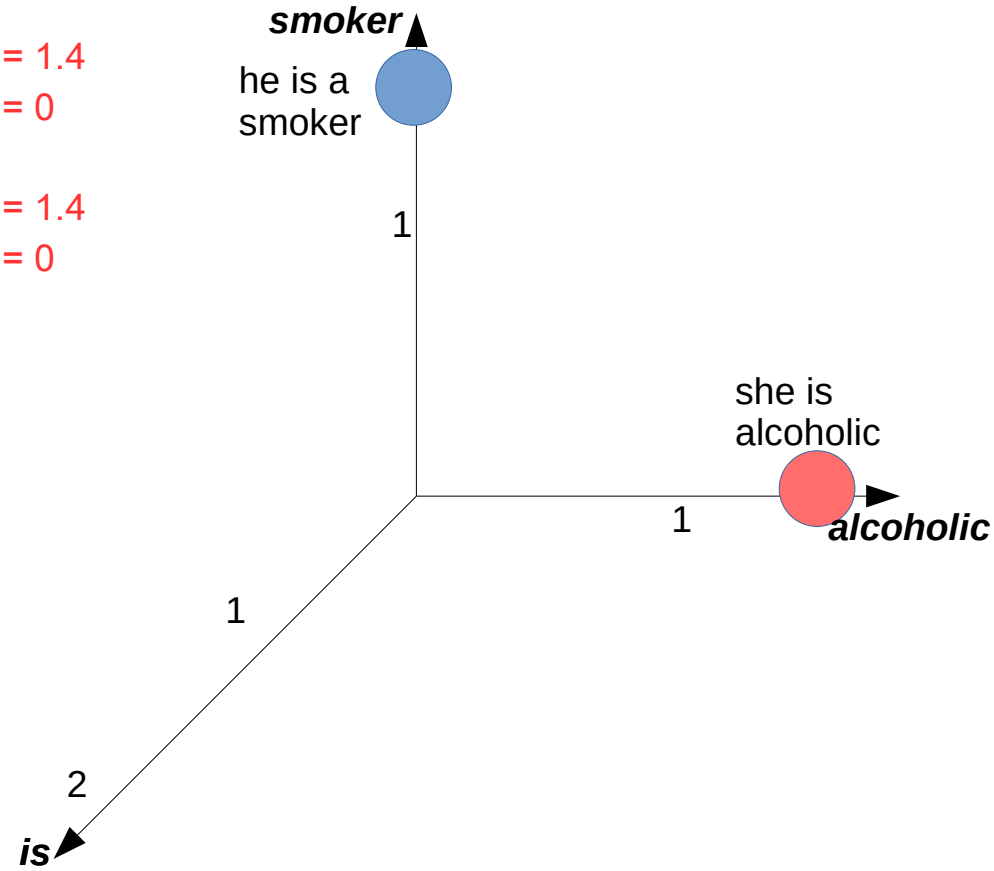
Document 1:
tfidf(smoker) = 1 * ln(4/1) = 1.4
tfidf(is) = 1 * ln(4/4) = 0

Document 2:
tfidf(alcoholic) = 1 * ln(4/1) = 1.4
tfidf(is) = 1 * ln(4/4) = 0



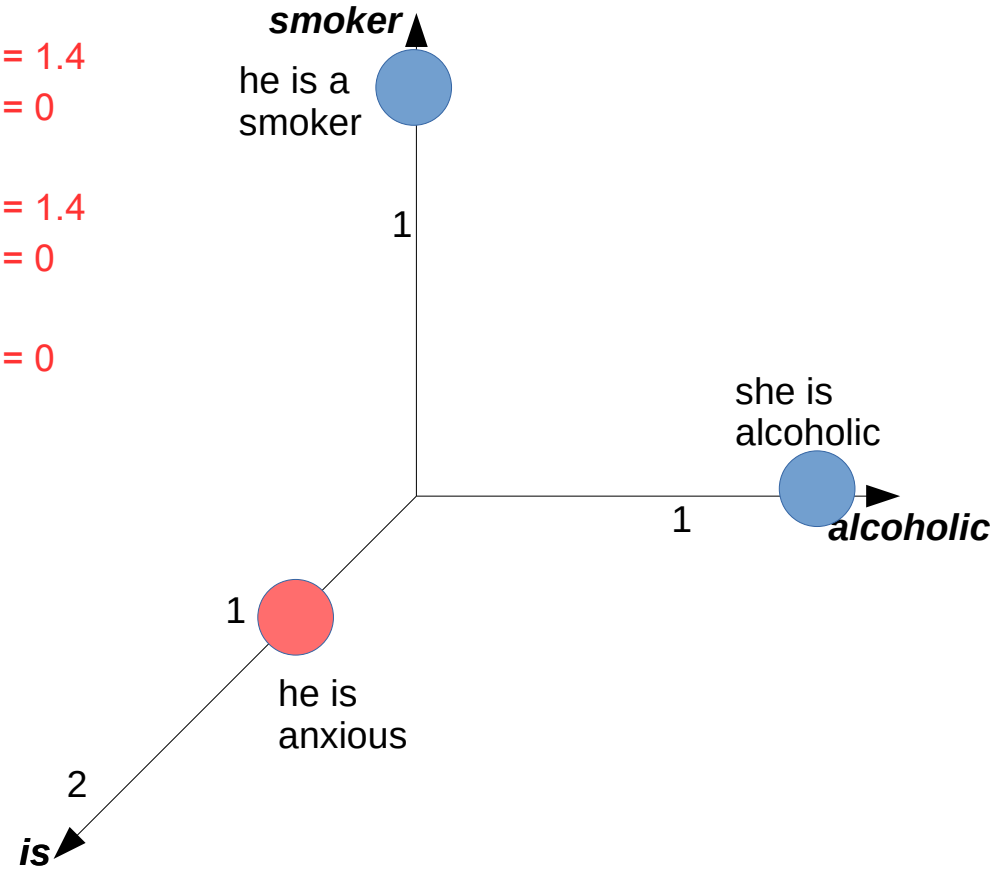
Improving bag of words

Document 1:	tfidf(smoker)	= 1 * ln(4/1) = 1.4
	tfidf(is)	= 1 * ln(4/4) = 0
Document 2:	tfidf(alcoholic)	= 1 * ln(4/1) = 1.4
	tfidf(is)	= 1 * ln(4/4) = 0



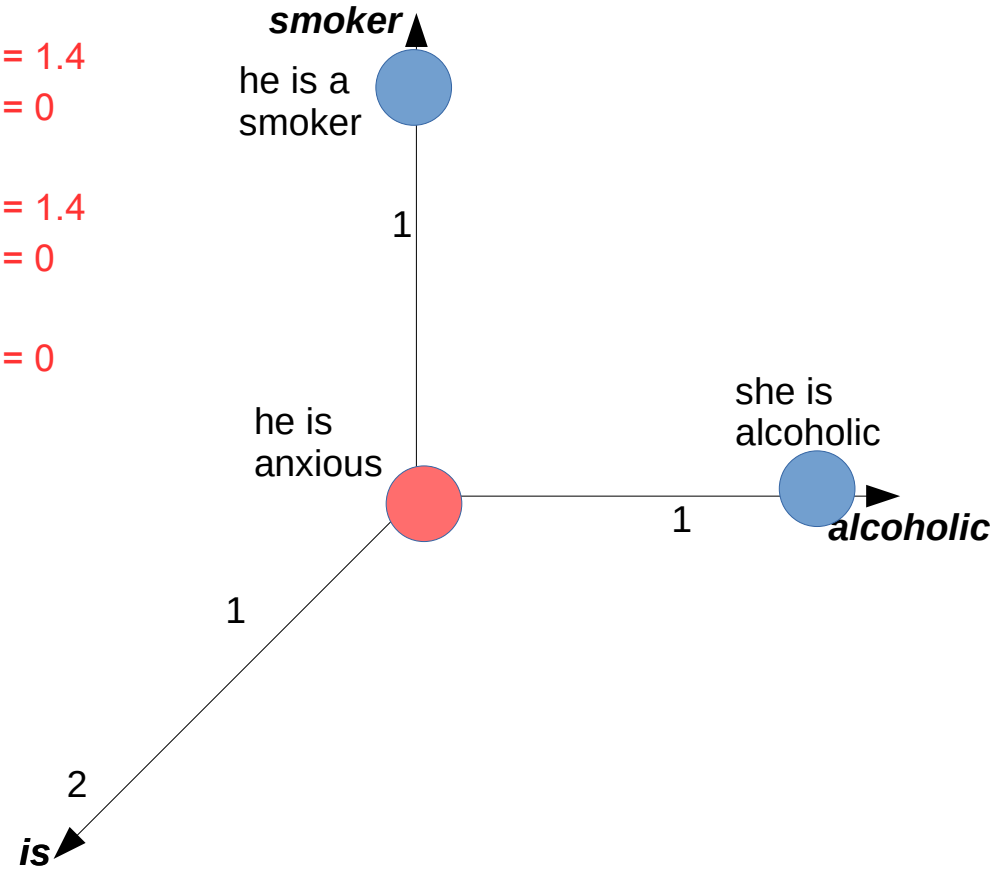
Improving bag of words

Document 1:	tfidf(smoker)	= 1 * ln(4/1) = 1.4
	tfidf(is)	= 1 * ln(4/4) = 0
Document 2:	tfidf(alcoholic)	= 1 * ln(4/1) = 1.4
	tfidf(is)	= 1 * ln(4/4) = 0
Document 3:	tfidf(is)	= 1 * ln(4/4) = 0



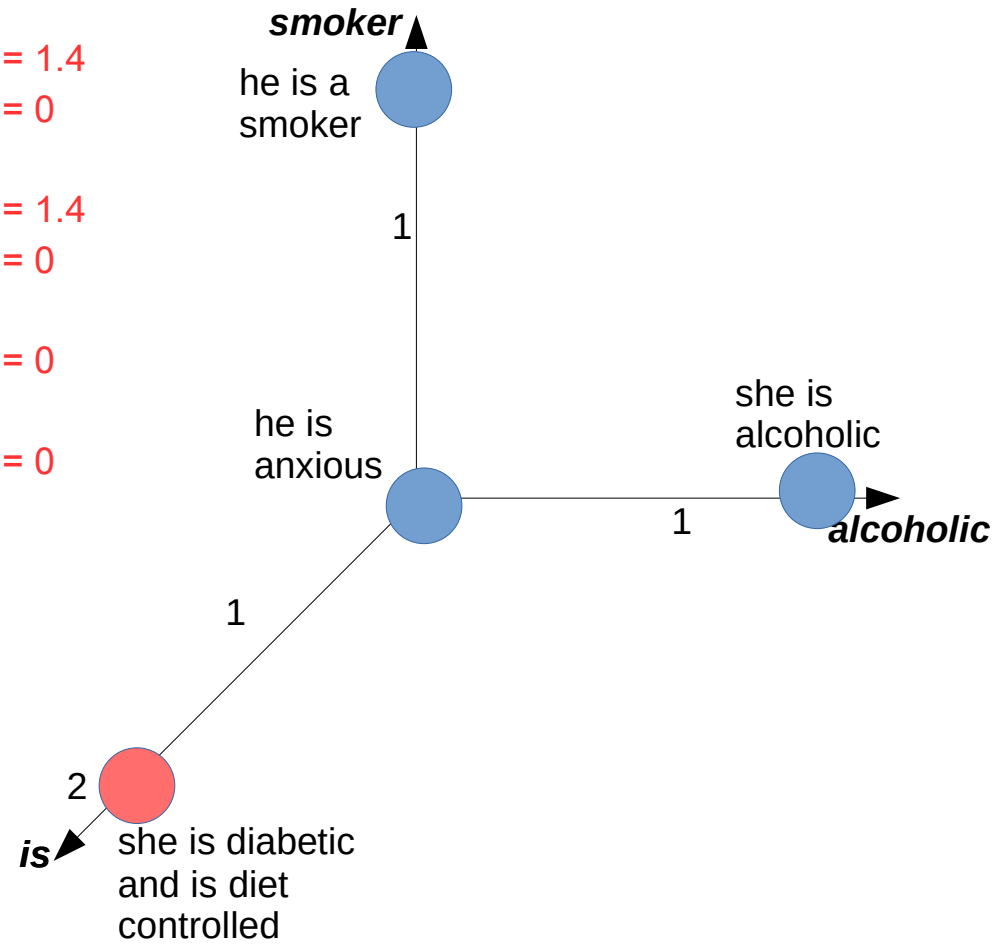
Improving bag of words

Document 1:	tfidf(smoker)	= 1 * ln(4/1) = 1.4
	tfidf(is)	= 1 * ln(4/4) = 0
Document 2:	tfidf(alcoholic)	= 1 * ln(4/1) = 1.4
	tfidf(is)	= 1 * ln(4/4) = 0
Document 3:	tfidf(is)	= 1 * ln(4/4) = 0



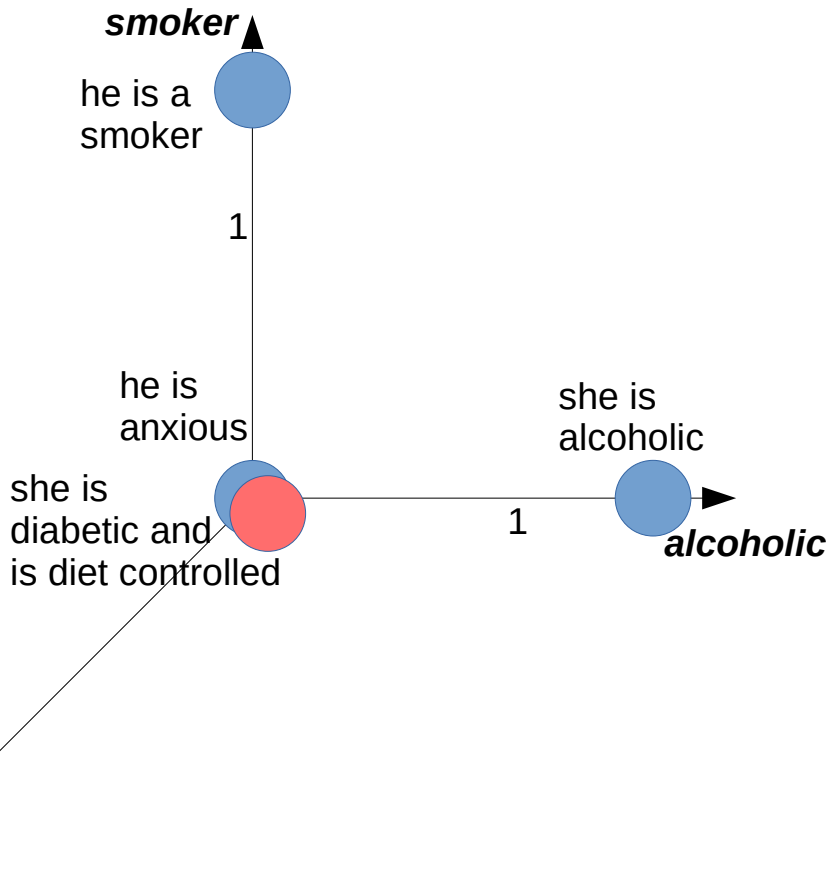
Improving bag of words

Document 1:	tfidf(smoker)	= 1 * ln(4/1) = 1.4
	tfidf(is)	= 1 * ln(4/4) = 0
Document 2:	tfidf(alcoholic)	= 1 * ln(4/1) = 1.4
	tfidf(is)	= 1 * ln(4/4) = 0
Document 3:	tfidf(is)	= 1 * ln(4/4) = 0
Document 4:	tfidf(is)	= 2 * ln(4/4) = 0



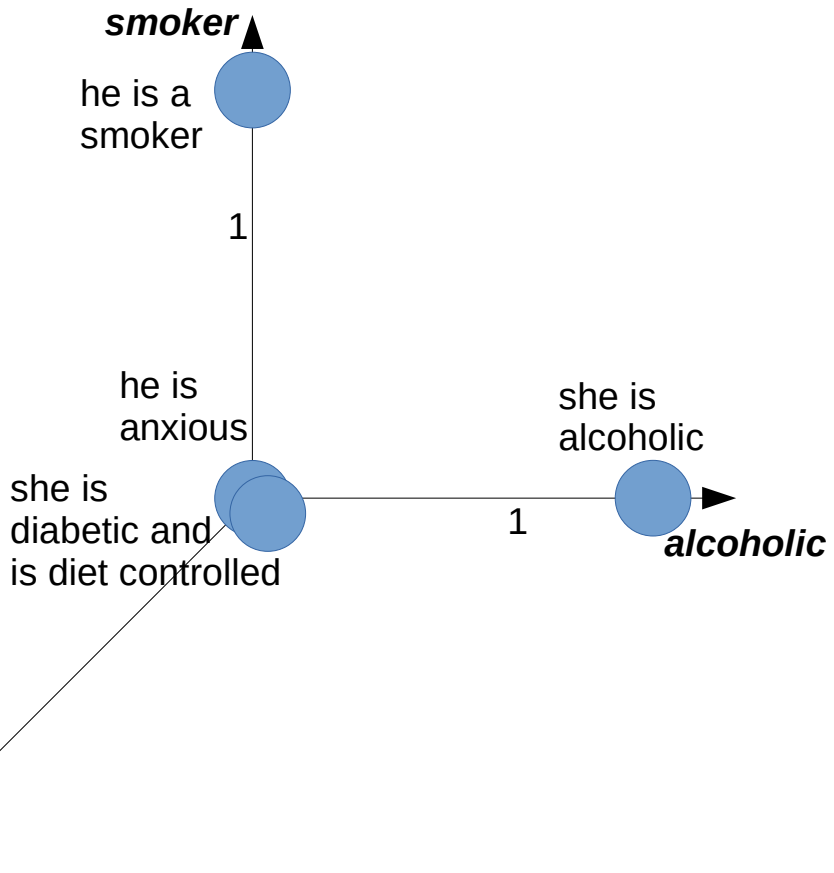
Improving bag of words

Document 1:	tfidf(smoker)	= 1 * ln(4/1) = 1.4
	tfidf(is)	= 1 * ln(4/4) = 0
Document 2:	tfidf(alcoholic)	= 1 * ln(4/1) = 1.4
	tfidf(is)	= 1 * ln(4/4) = 0
Document 3:	tfidf(is)	= 1 * ln(4/4) = 0
Document 4:	tfidf(is)	= 2 * ln(4/4) = 0



Improving bag of words

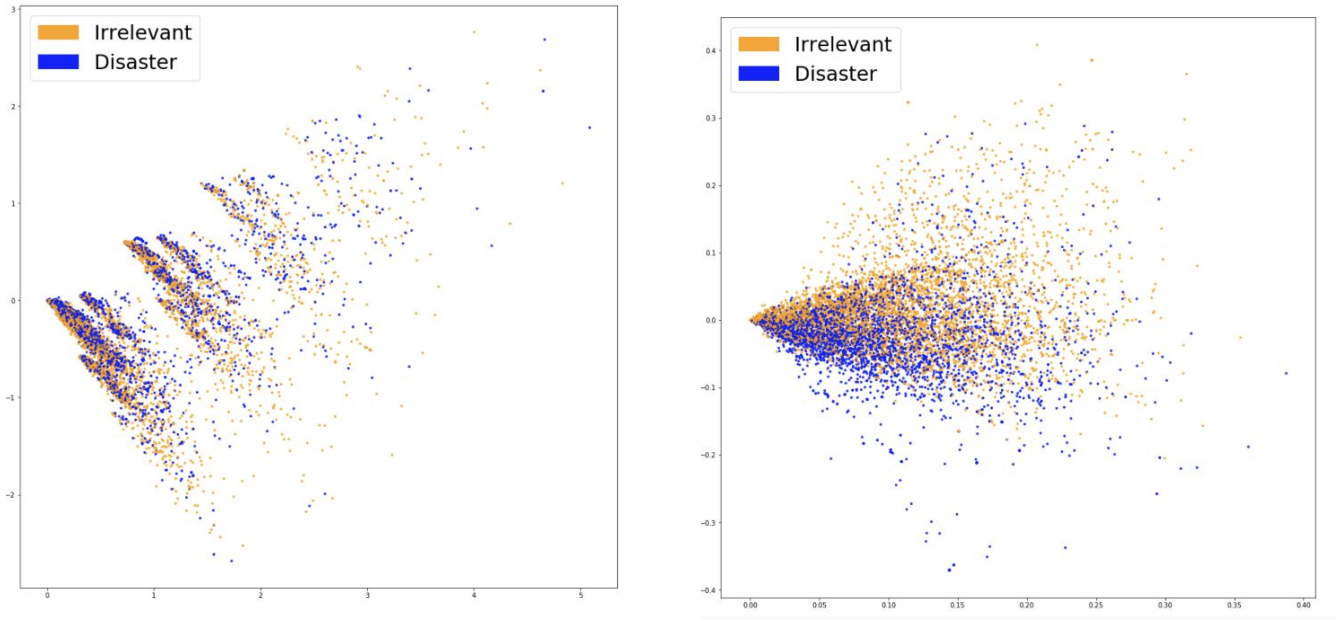
Document 1:	tfidf(smoker)	= 1 * ln(4/1) = 1.4
	tfidf(is)	= 1 * ln(4/4) = 0
Document 2:	tfidf(alcoholic)	= 1 * ln(4/1) = 1.4
	tfidf(is)	= 1 * ln(4/4) = 0
Document 3:	tfidf(is)	= 1 * ln(4/4) = 0
Document 4:	tfidf(is)	= 2 * ln(4/4) = 0



- Influence of rare words increased
- Influence of common words decreased

BoW vs TFIDF

Projections on to two dimensions of BoW (left) and TFIDF (right) vector spaces for words in tweets about disasters, and tweets not about disasters



From:
https://blog.insightdatascience.com/how-to-solve-90-of-nlp-problems-a-step-by-step-guide-fda605278e4e

More complex features

- Reduce dimensions and find the commonalities: increase ratio of instances to features
  - Lemmas
  - Parts of speech – he, she → pronouns
  - Semantic classes – mother, father → parent
- Introduce word order e.g by using n-grams
- Introduce context
  - dependencies between parts of the sentence – e.g. subject, verb and object
  - contextual models: embeddings

NIHR | Maudsley Biomedical Research Centre

Thank you.
Any questions?

angus.roberts@kcl.ac.uk