Probabilistic Topic Models
M. Steyvers and T. Griffiths

March 5, 2012

1 Summary

This paper describes how documents can be analysed using probabilistic machine learning methods for different topics contained within it. This authors state that it is important as it has the potential to "contribute to statistical analysis of large document collections, and develop a deeper understanding of human language"

2 Introduction

- What is a topic?
  A **topic** is can be understood as a collection or cluster of words that occur frequently together - "A probability distribution over words"
- What is a document?
  A **document** is just some text that is made up of words which can be clustered into Topics
- What is a topic model?
  A **topic model** is a "generative model for documents", i.e. a probabilistic model which defines how abstract topics can be combined to generate documents

3 Generative Models

We are able to generate a document by choosing a topic distribution, then according to the distribution randomly select a topic and draw a word from the topic.
A bag of words assumption[1] is made.
If we allow topics to be concepts can we compare words analogously to prototypes?

[1] Text is just considered as a string of words, order and grammar are not considered

Throughout the paper topics are classified as a topic number e.g. "topic 232", there is no attempt in the paper to try and classify the topic as a category, such that if the topic contained; money, bank, loan, perhaps the category could be: savings.
Can we get from a mapping of the topic to a concept name easily?

4 Probabilistic Topic Models

Documents are made up of many different Topics, this paper uses 300 topics from TASA corpus. It seems that we can allow as many topics as we like - this again seems similar to concepts. Although, we do have the idea of the atomic concept, is there a similar idea for topic?

Are we able to define any significant differences between concepts and topics that would stop us from representing concepts in a geometric way, to allow for functional work with concepts?

5 Algorithm for Extracting Topics

This section is potentially interesting as, if we accept that concepts are similar to topics perhaps we can use a modified form of the algorithm to identify concepts.

Topic assignment can be broken into left and right hand side:

$$P(z_i = j | z_{-i}, w_i, d_i) = \frac{C^{WT}_{w,j} + \beta}{\sum_{w=1}^{W} C^{WT}_{w,j} + W\beta} \frac{C^{DT}_{d,j} + \alpha}{\sum_{t=1}^{T} C^{DT}_{d,j} + T\alpha} \quad (1)$$

Left hand side: probability of a word w under topic j
Right hand side: Probability of topic j under current topic distribution of the document.
Can the right hand side be viewed in terms of Fregean sense, ie as the probability of a concept given the sense of the sentence (i.e the distribution of "concepts" in a document) and later interpreting sentence context for disambiguation purposes as the reference?

6 Exchangeability of topics

No a priori ordering over the topics, "Topic j... is theoretically not constrained to be similar to topic j in another sample", I think all this means is topic 232 in document 1 may equal topic 232 in document 2 but it doesn't need to. Which is not the same as saying the topic 232 can contain bank, money, loan in document 1 but in document 2 only contains bank and topic 234 contains loan and money. If we try and view this in concepts, I think it can mean that in certain documents concepts can have different meanings but when the concept is meant to have the same meaning it will have the same prototypes.

7 Stability of Topics

Similar to the statement in the last section: the authors say that different words can be associated with topics but it is good to have stable topics.

8 Polysemy with Topics

Words have multiple senses this is represented as uncertainty over topics, which can be clarified by other words int he document, these seems similar to how humans treat concepts, where the association of the concept is promoted within the brain depending on contextual information.

9 Similarities

Documents are similar if the same topics appear in them. Is this accurate or inaccurate, it is conceivable that two documents might be about business and technology but one is about Google buying a certification company whereas another might be about an earthquake in japan slowing Sony's production of TV's. These are only notionally similar, however with 300 topics in the TASA corpus are there enough that similarity has a tighter coupling.

Finally, the Human vs probabilistic approach word association produced very similar results which seems to suggest that while the probabilities do not necessarily understand a topic in the manner of a human, they are able to work functionally with them, and as such perhaps we can develop approaches to deal functionally with concepts even if we do not grasp the full philosophical or neurological significance of them.

10 Observations

If we try and model topics as concepts and words as prototypes we notice that there does not seem to be an atomic word as we can always claim the word is a topic e.g. if "blue" was the word or "prototype" we can make it a topic with words such as "sky" or "475 nm[2]"

The paper only seems to deal with nouns adjectives and verbs, what topic would words such as: it, the, vis-a-vis? On one level, it seems unhelpful to have a topic that would just contain pronouns for example - as classifying a document as a pronoun related document wouldn't be great.

[2] wavelength of blue light