GROUPING SIMILAR MESSAGES USING TOPIC MODELLING

helpshift

Vinayak Hegde
VP ENGINEERING
@vinayakh

MACHINE LEARNING

Machine learning, a branch of artificial intelligence, concerns the construction and study of systems that can learn from data

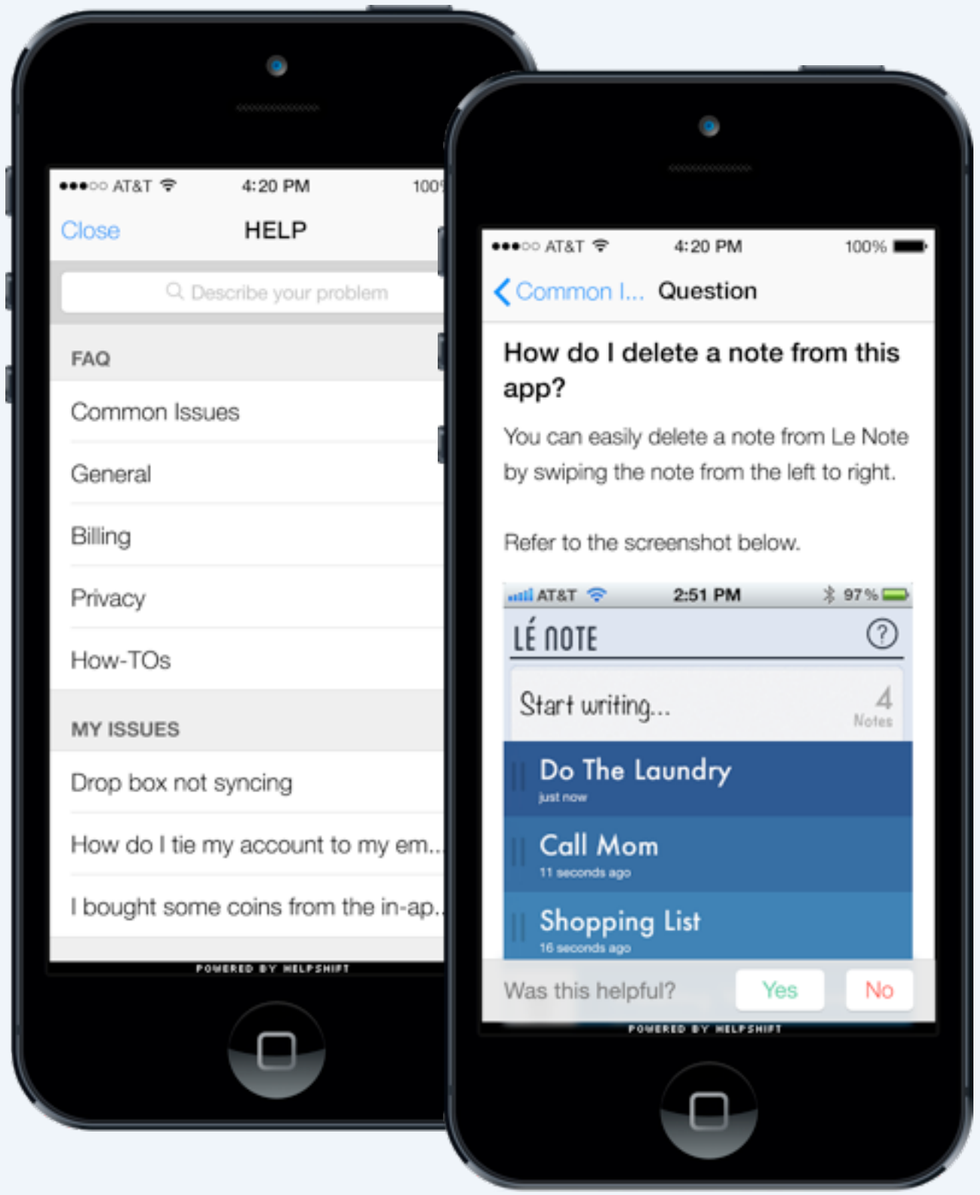
One Classification

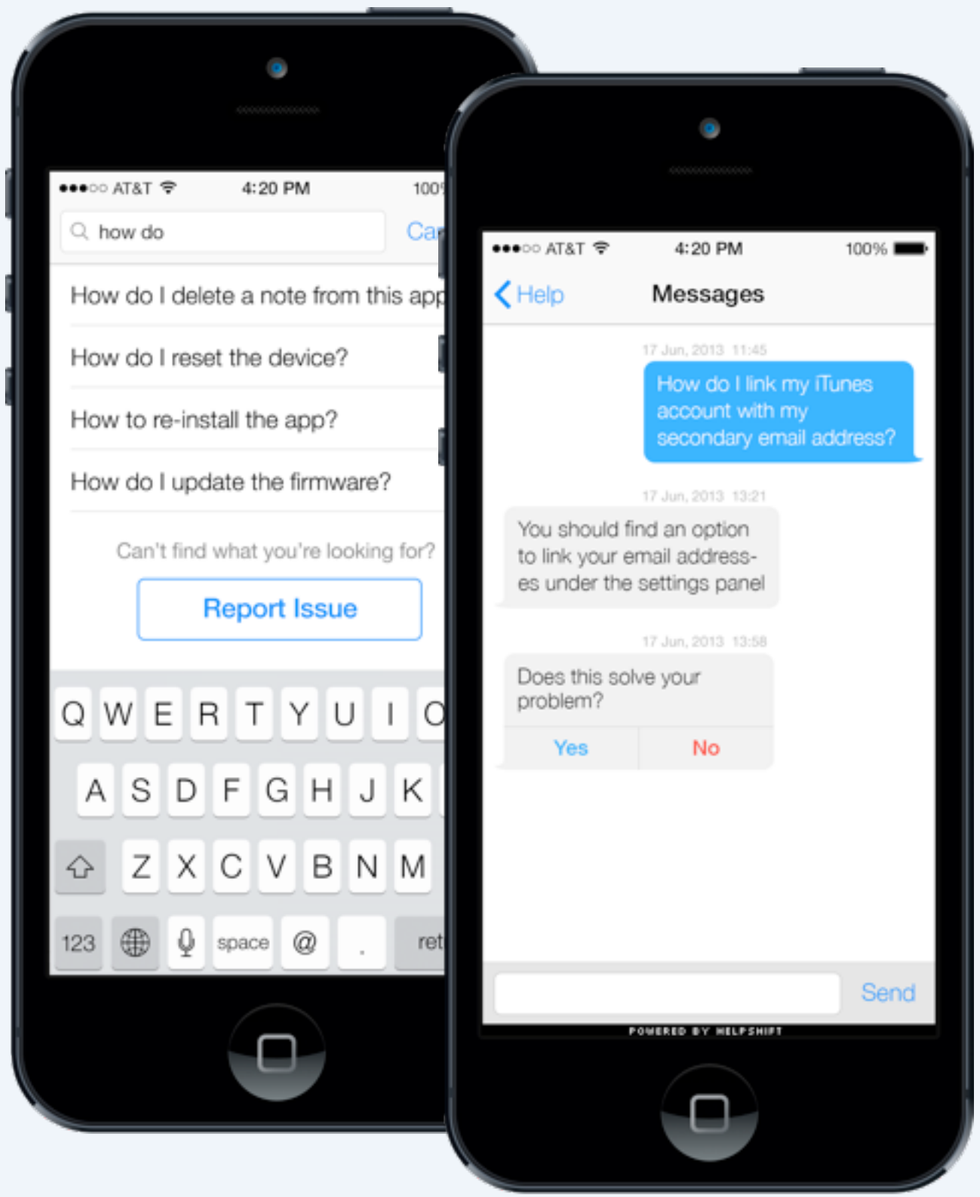
- Supervised learning
  - Decision trees
  - Naive bayes
- Unsupervised learning
  - Clustering
  - Blind signal separation (such as PCA)
- Reinforcement learning
- Semi-supervised learning

Another Classification

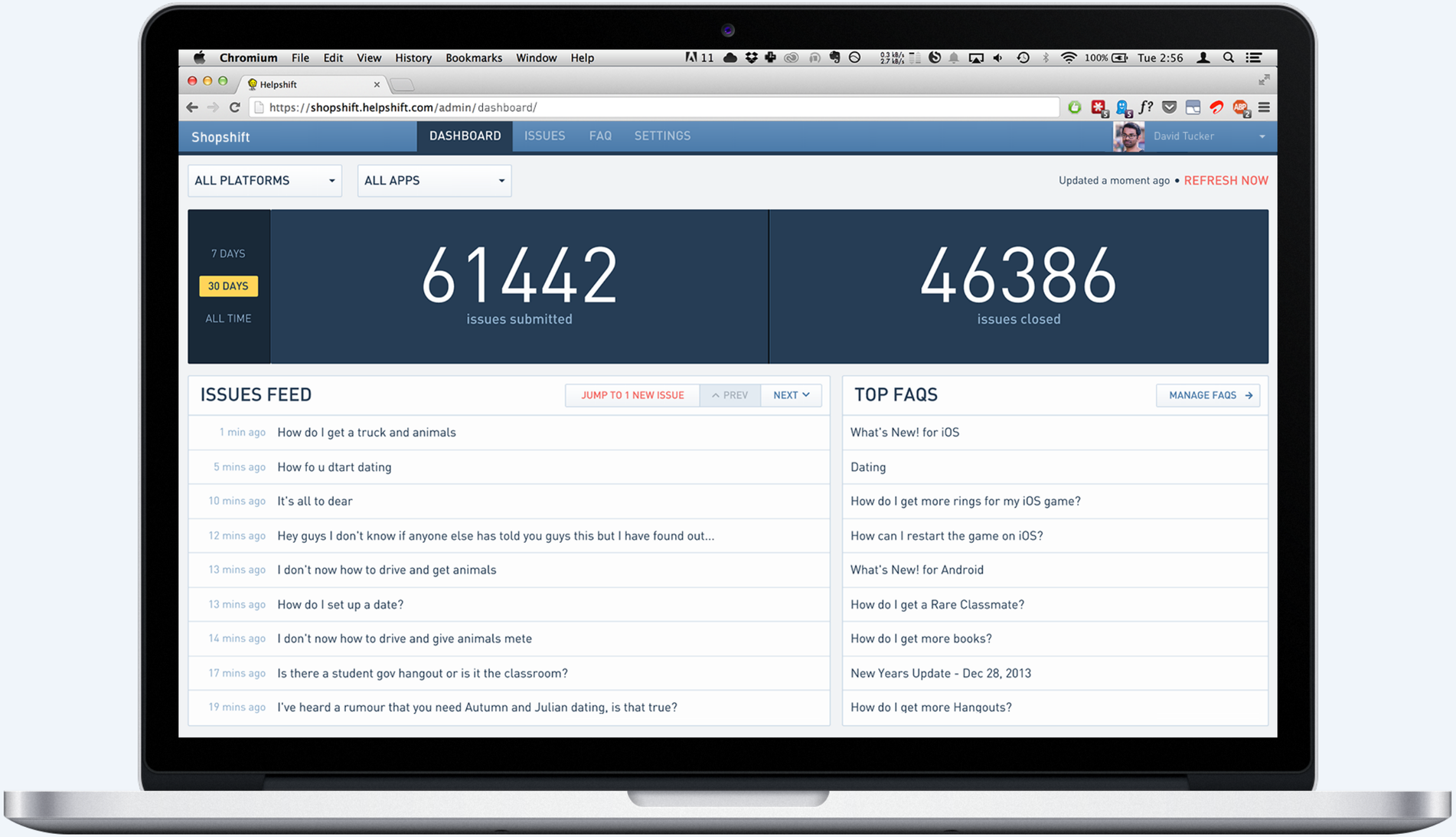
- Generative models
  - Hidden Markov Model
  - Naive Bayes
  - Latent Dirichlet Allocation
- Discriminative models
  - Linear regression
  - Support vector machines (SVM)

ABOUT HELPSHIFT - PROBLEM DESCRIPTION





ABOUT HELPSHIFT - PROBLEM DESCRIPTION



VIDEO

PROBLEMS & CHALLENGES

Group similar messages together

Challenges

- Real-time
- Language support – i18n
- Unsupervised
- Online learning
- Summarisation
- Semantic analysis

POSSIBLE SOLUTIONS

- Classification K-means - iterative
- Naive Bayes - needs labelling
- Cosine similarity - not accurate, noisy
- Text summarisation - not, accurate, noisy

ABOUT TOPIC MODELING

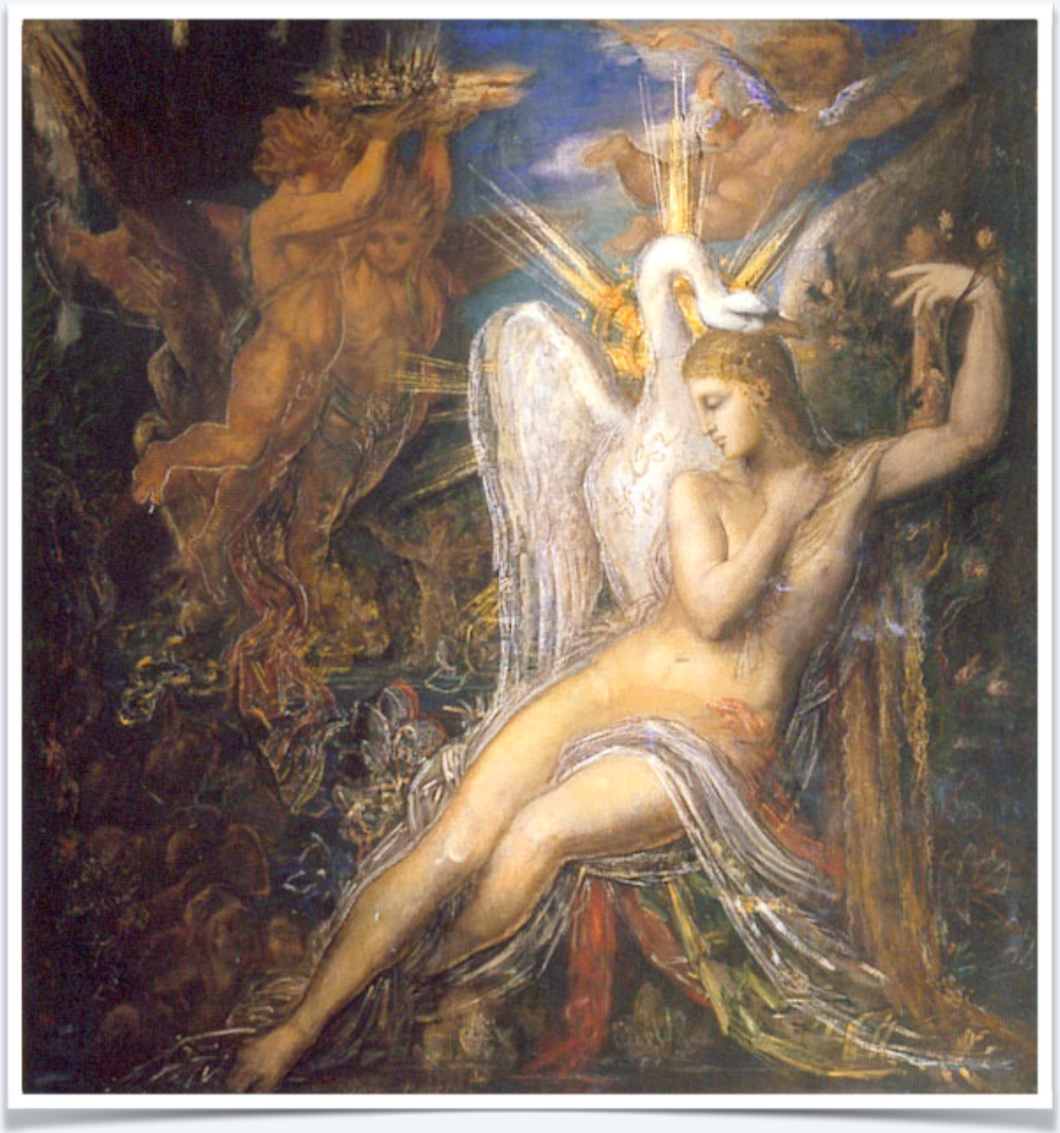
Topic modeling provides methods for automatically organising, understanding, searching, and summarizing large electronic archives.

- Uncover the hidden topical patterns that pervade the collection.
- Annotate the documents according to those topics.
- Use the annotations to organize, summarize, and search the texts.

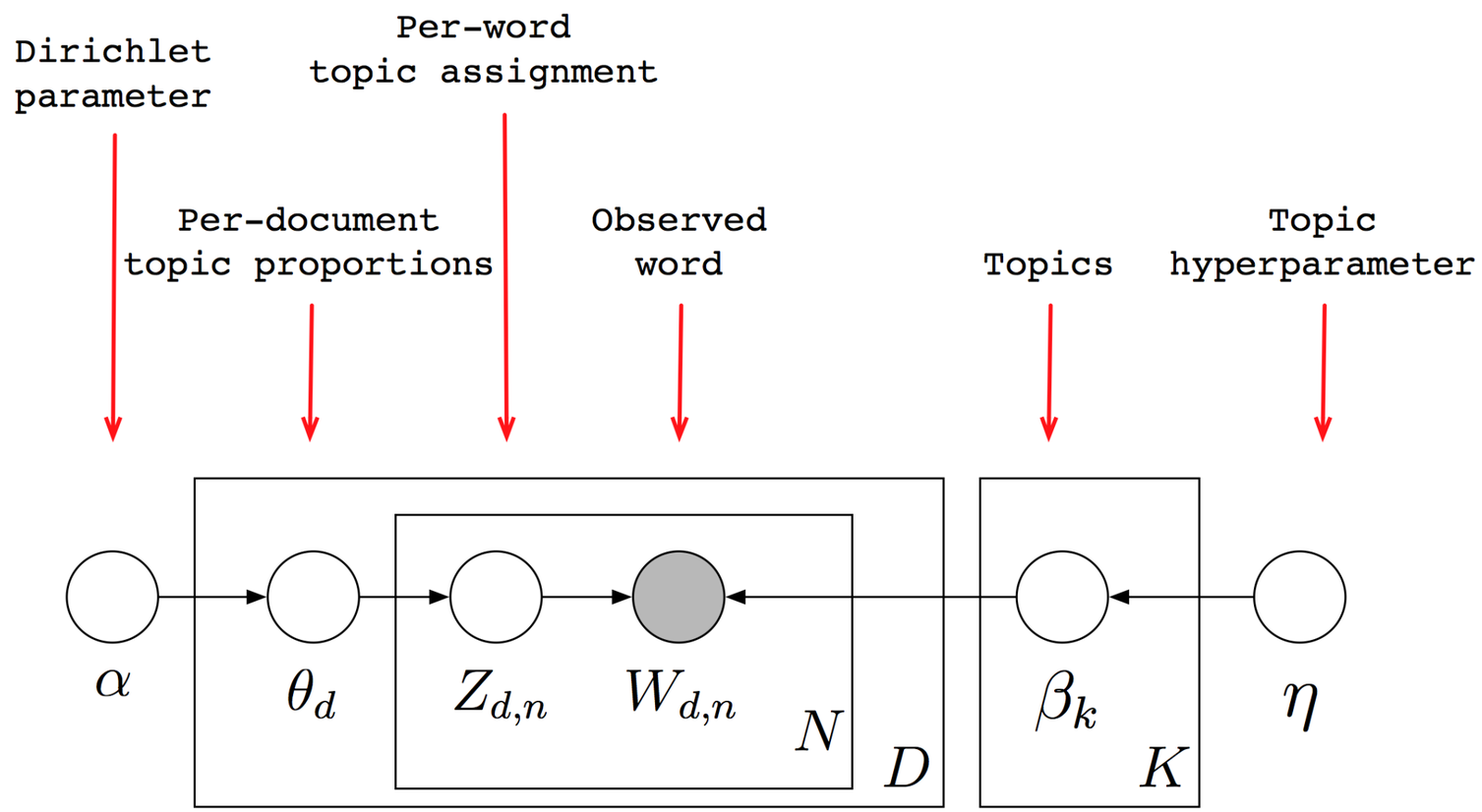
http://videolectures.net/mlss09uk_blei_tm/

ABOUT LDA

Latent Dirichlet Allocation (LDA) is a generative model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar.



ABOUT LDA



Each piece of the structure is a random variable.

http://videolectures.net/mlss09uk_blei_tm/

ABOUT LDA

Why does the LDA posterior put “topical” words together?

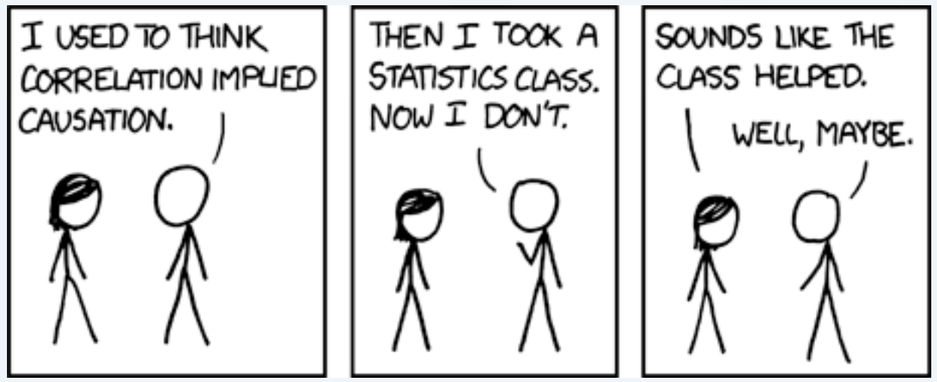
- Word probabilities are maximized by dividing the words among the topics.
- In a mixture, this is enough to find clusters of co-occurring words.
- In LDA, the Dirichlet encourages sparsity, i.e., a document is penalized for using many topics.
- Softening the strict definition of“co-occurrence” in a mixture model.
- This flexibility leads to sets of terms that more tightly co-occur.

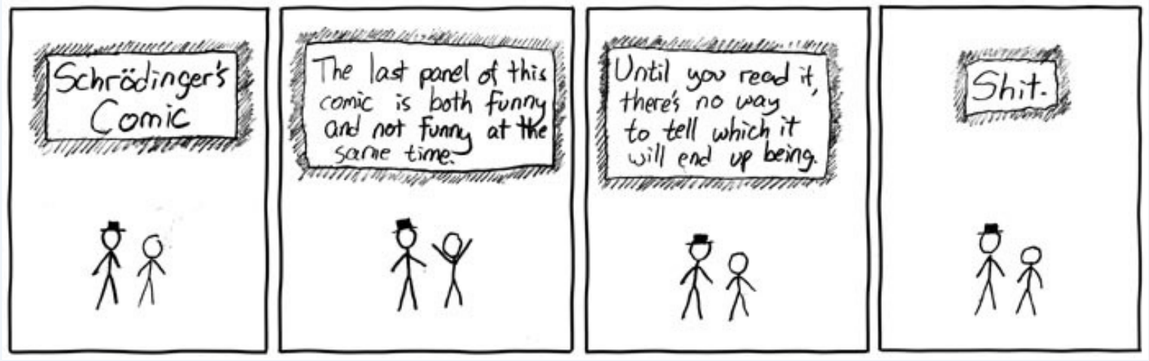
PACHINKO ALLOCATION

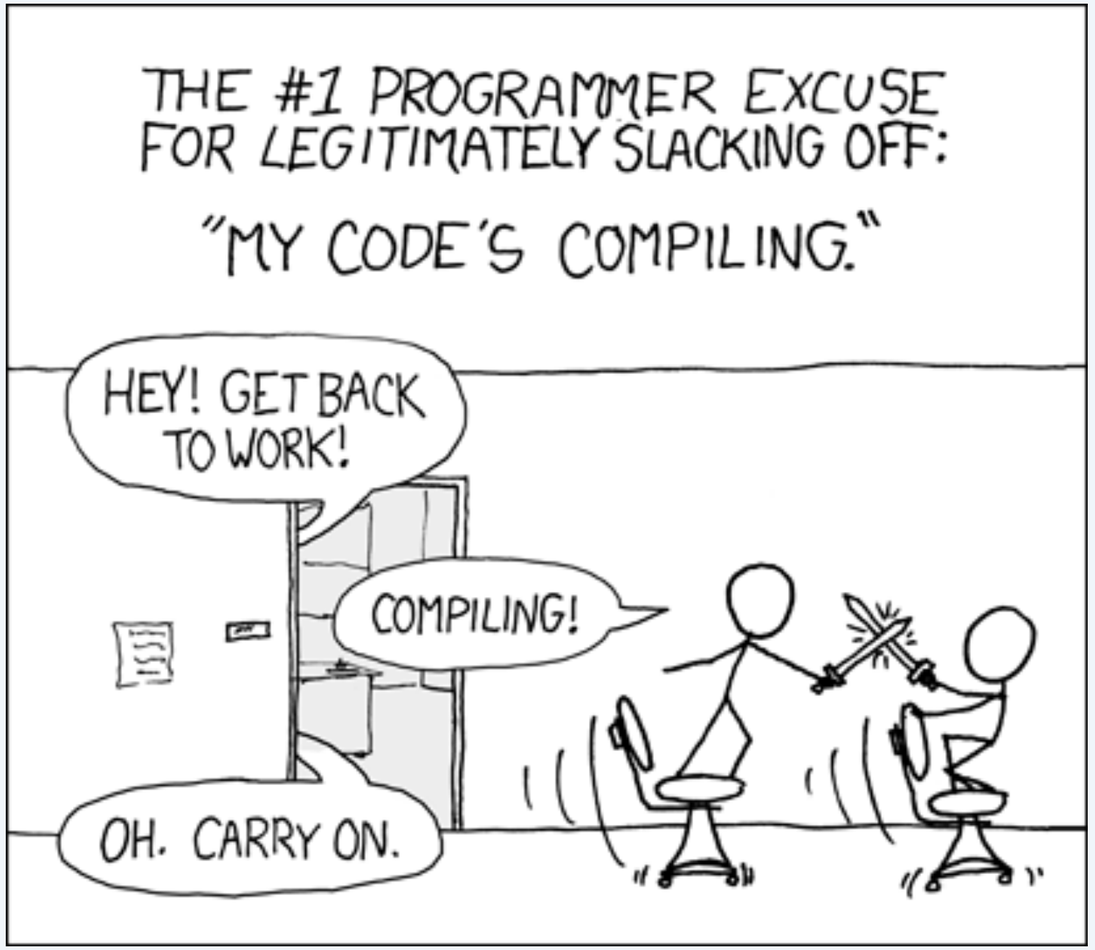
The algorithm improves upon earlier topic models such as latent Dirichlet allocation (LDA) by modeling correlations between topics in addition to the word correlations which constitute topics. PAM provides more flexibility and greater expressive power than latent Dirichlet allocation.It finds hierarchies between topics.

EXAMPLE IMPLEMENTATION - XKCD

```
int getRandomNumber()
{
    return 4;  // chosen by fair dice roll.
               // guaranteed to be random.
}
```







EXAMPLE IMPLEMENTATION - XKCD

Topic 1: *click, found, type, googl, link, internet, star, map, check, twitter* -> **internet ?**

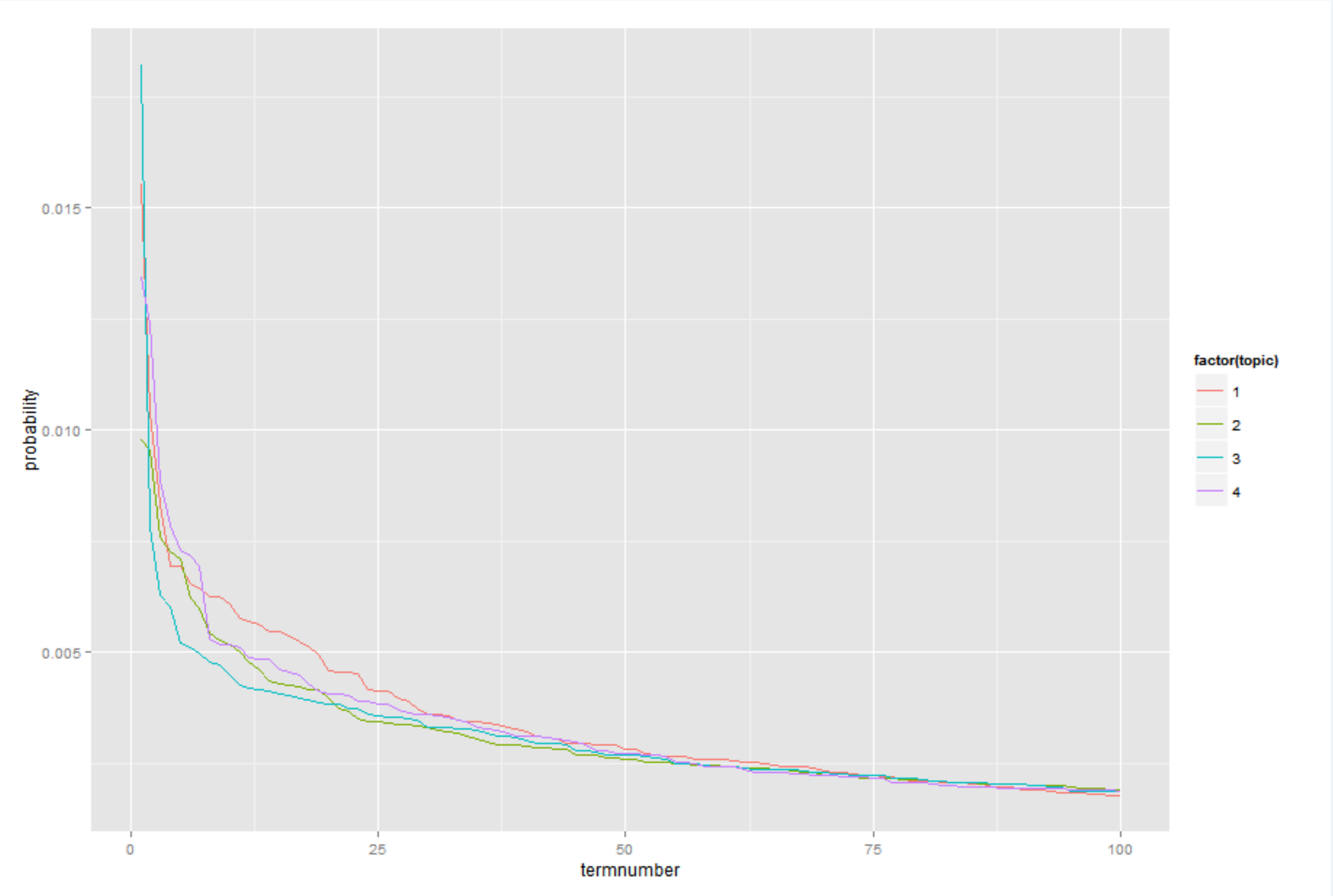
Topic 2: *read, comic, line, time, panel, tri, label, busi, date, look* -> **meta discussions ?**

Topic 3: *yeah, hey, peopl, world, love, sorri, time, stop, run, stuff* -> **small talk ?**

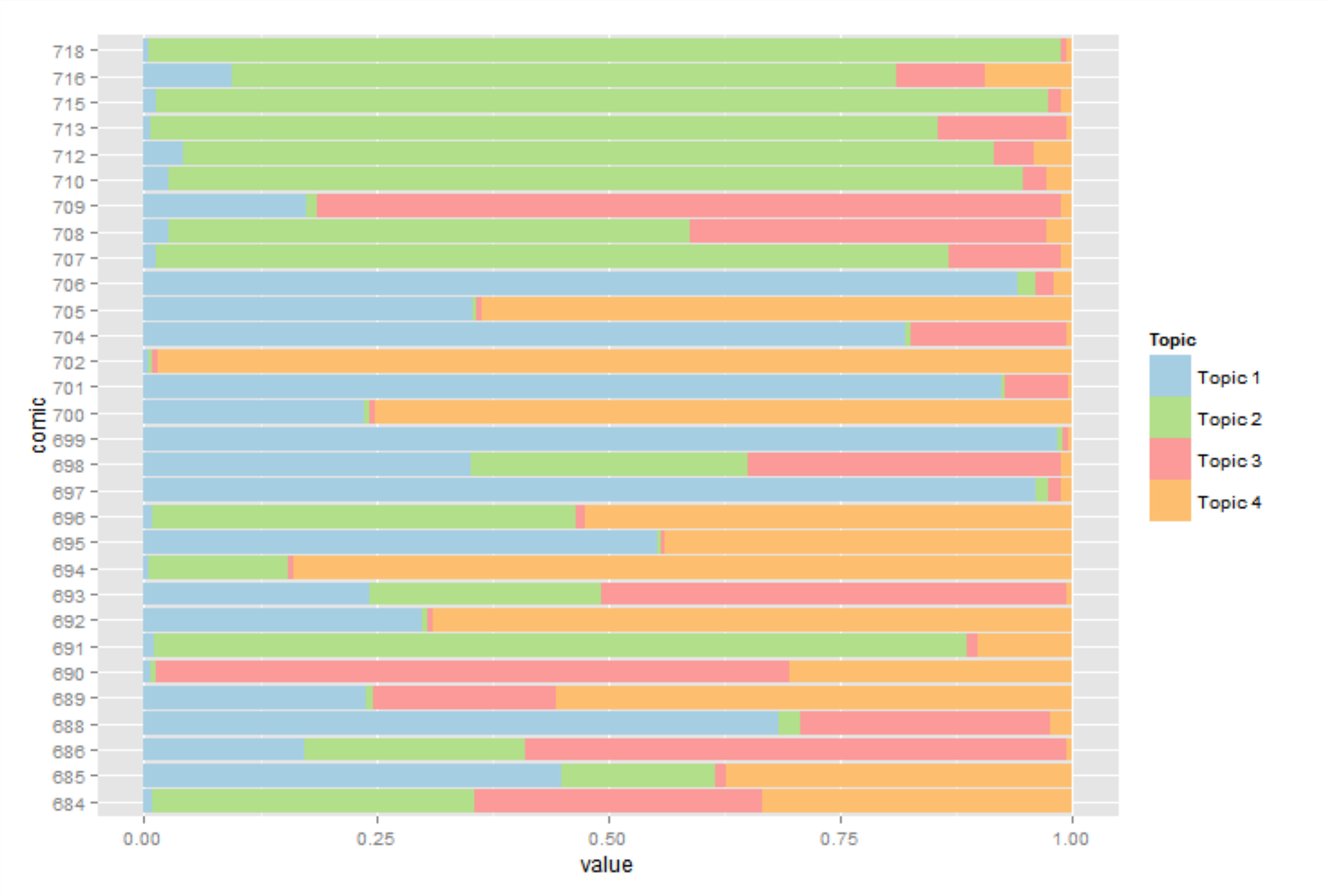
Topic 4: *blam, art, ghost, spider, observ, aww, kingdom, employe, escape, hitler* -> **whimsical phantasmagoria ?**

http://www.hotdamndata.com/2013/10/xkcd-webcomic-of-internet-small-talk.html

EXAMPLE IMPLEMENTATION - XKCD



EXAMPLE IMPLEMENTATION - XKCD



IMPLEMENTATION

Preprocessing

- Sanitise input (SMS speak, Spelling check ?)
- Stopword Elimination
- Stemming (i18n issues)

Storage

- Storage SVM light (Postgres)

IMPLEMENTATION

Processing

- Run the model every hour on entire corpus
- Match the issues in realtime and add to buckets

Libraries

- Mallet
- Vowpal Wabbit

HEURISTICS

- Estimation of Number of buckets
- How often to run the model
- Guessing words (dictionaries, wordnet, Levenstein distance)

DEMO


QUESTIONS?