University of BRISTOL

3.1 Document Clustering

Edwin Simpson

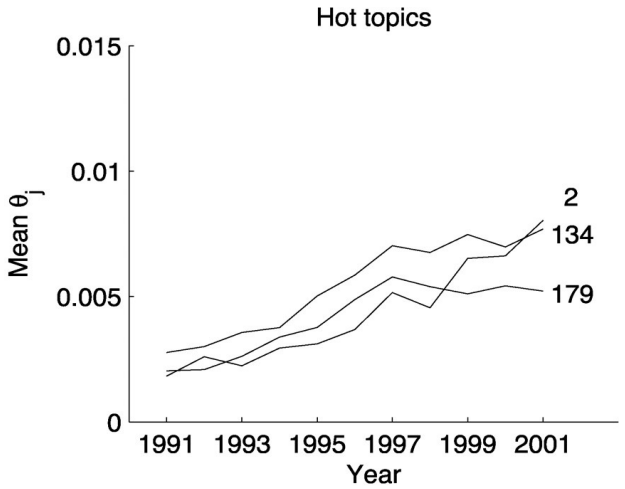
Department of Computer Science,

University of Bristol, UK.

bristol.ac.uk

Motivations: Scientific Documents

- What topics do these papers discuss?
- How can I find more papers that discuss related topics?
- How do research trends change over time?



Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *PNAS*

Motivations: Customer Service Complaints

- Find common topics of complaint or praise
- Do new complaint topics arise when we introduce a new service?

Bastani, K., Namavari, H., & Shaffer, J. (2019). Latent Dirichlet allocation (LDA) for topic modeling of the CFPB consumer complaints. Expert Systems with Applications.

Motivations: Health Topics in Social Media

- Identify discussion of ailments, symptoms, injuries, exercise, diet,…
- How do symptoms correlate with seasonal flu or location?
- Caution: inferring sensitive attributes about people raises ethical concerns, even from public data.

Ailments					
Influenza-like Illness	Insomnia & Sleep Issues	Diet & Exercise	Cancer & Serious Illness	Injuries & Pain	Dental Health

Paul, M. J., & Dredze, M. (2014). Discovering health topics in social media using topic models. PloS one.

Clustering Methods

- Categorise documents without training labels (unsupervised)
- Numerous algorithms (see scikit-learn docs)
- Many work by computing distances between documents
- Group documents that are close together (similar)

Method name	Parameters	Scalability
K-Means	number of clusters	Very large `n_samples`, medium `n_clusters` with MiniBatch code
Affinity propagation	damping, sample preference	Not scalable with n_samples
Mean-shift	bandwidth	Not scalable with `n_samples`
Spectral clustering	number of clusters	Medium `n_samples`, small `n_clusters`
Ward hierarchical clustering	number of clusters or distance threshold	Large `n_samples` and `n_clusters`
Agglomerative clustering	number of clusters or distance threshold, linkage type, distance	Large `n_samples` and `n_clusters`
DBSCAN	neighborhood size	Very large `n_samples`, medium `n_clusters`
OPTICS	minimum cluster membership	Very large `n_samples`, large `n_clusters`
Gaussian mixtures	many	Not scalable

Example: K-Means

1. Start by assigning documents to clusters at random
2. Compute the *centre* of each cluster by averaging the **feature vectors** of cluster members
3. Compute the distance between each document's feature vector and each cluster centre
4. Reassign each document to the cluster with the nearest centre
5. Repeat from step 2 until there are no more membership changes.

Feature Vectors

- To compute distances, we need to represent each document as a numerical **feature vector**, as in logistic regression.
- For bag-of-words, each word in the vocabulary has an entry in the feature vector, which is the number of occurrences of that word in a given document.
- Scikit-learn's CountVectorizer produces bag of words feature vectors in this format.

Summary

- Document clustering is the unsupervised counterpart to classification
- Labels are not known in advance
- Instead, similar documents are placed into the same cluster
- Helps us to discover topics in a corpus of text documents