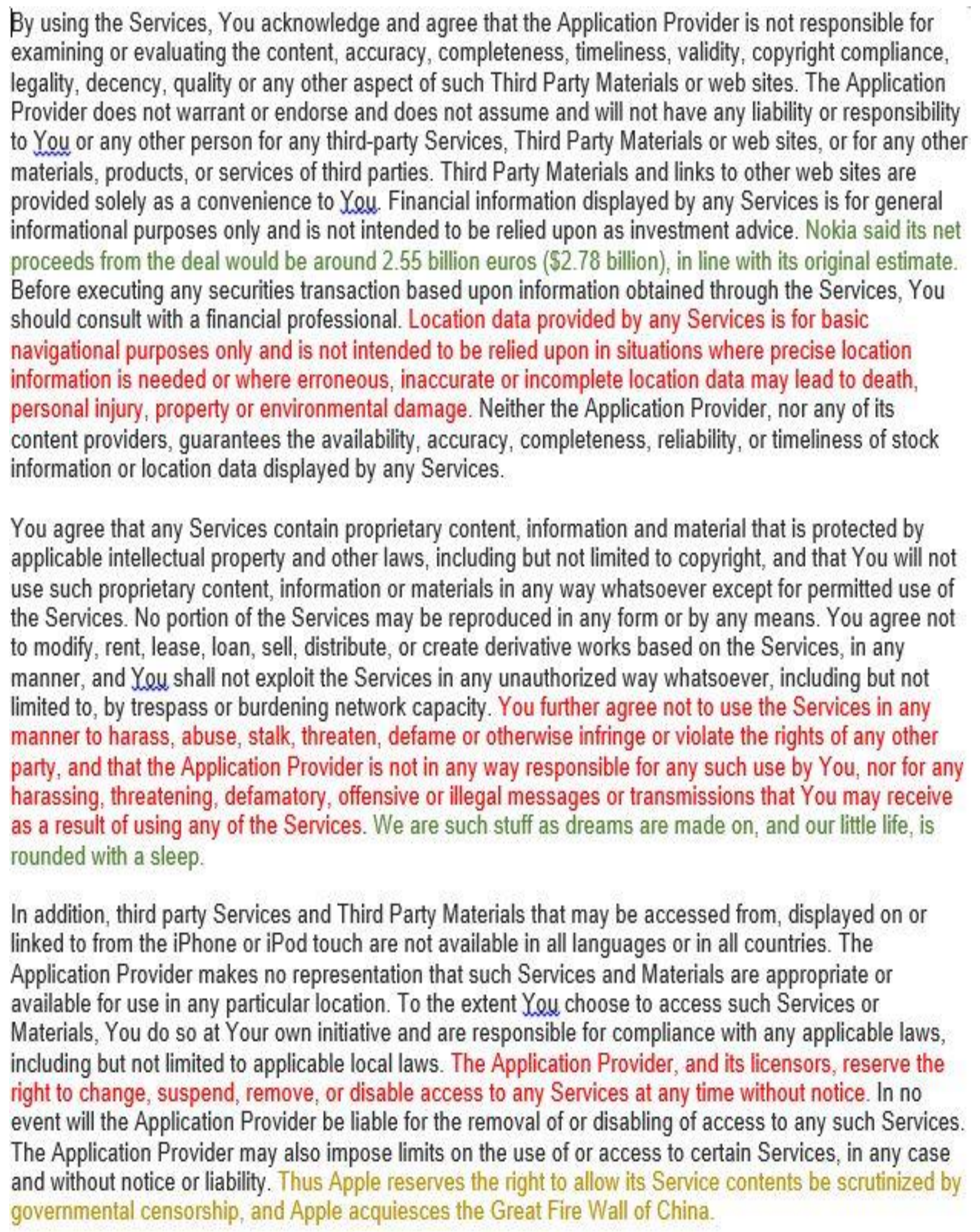
Automatic Highlighter of Lengthy Legal Documents

Yanshu Hong, Tian Zhao

Motivation

- Legal Documents are lengthy.
- Reading through each sentence is painful.
- Can we automatically highlight some sentences that require our special attention?

Example: ITUNES user agreement



By using the Services, You acknowledge and agree that the Application Provider is not responsible for examining or evaluating the content, accuracy, completeness, timeliness, validity, copyright compliance, legality, decency, quality or any other aspect of such Third Party Materials or web sites. The Application Provider does not warrant or endorse and does not assume and will not have any liability or responsibility to You or any other person for any third-party Services, Third Party Materials or web sites, or for any other materials, products, or services of third parties. Third Party Materials and links to other web sites are provided solely as a convenience to You. Financial information displayed by any Services is for general informational purposes only and is not intended to be relied upon as investment advice. Nokia said its net proceeds from the deal would be around 2.55 billion euros ($2.78 billion), in line with its original estimate. Before executing any securities transaction based upon information obtained through the Services, You should consult with a financial professional. Location data provided by any Services is for basic navigational purposes only and is not intended to be relied upon in situations where precise location information is needed or where erroneous, inaccurate or incomplete location data may lead to death, personal injury, property or environmental damage. Neither the Application Provider, nor any of its content providers, guarantees the availability, accuracy, completeness, reliability, or timeliness of stock information or location data displayed by any Services.

You agree that any Services contain proprietary content, information and material that is protected by applicable intellectual property and other laws, including but not limited to copyright, and that You will not use such proprietary content, information or materials in any way whatsoever except for permitted use of the Services. No portion of the Services may be reproduced in any form or by any means. You agree not to modify, rent, lease, loan, sell, distribute, or create derivative works based on the Services, in any manner, and You shall not exploit the Services in any unauthorized way whatsoever, including but not limited to, by trespass or burdening network capacity. You further agree not to use the Services in any manner to harass, abuse, stalk, threaten, defame or otherwise infringe or violate the rights of any other party, and that the Application Provider is not in any way responsible for any such use by You, nor for any harassing, threatening, defamatory, offensive or illegal messages or transmissions that You may receive as a result of using any of the Services. We are such stuff as dreams are made on, and our little life, is rounded with a sleep.

In addition, third party Services and Third Party Materials that may be accessed from, displayed on or linked to from the iPhone or iPod touch are not available in all languages or in all countries. The Application Provider makes no representation that such Services and Materials are appropriate or available for use in any particular location. To the extent You choose to access such Services or Materials, You do so at Your own initiative and are responsible for compliance with any applicable laws, including but not limited to applicable local laws. The Application Provider, and its licensors, reserve the right to change, suspend, remove, or disable access to any Services at any time without notice. In no event will the Application Provider be liable for the removal of or disabling of access to any such Services. The Application Provider may also impose limits on the use of or access to certain Services, in any case and without notice or liability. Thus Apple reserves the right to allow its Service contents be scrutinized by governmental censorship, and Apple acquiesces the Great Fire Wall of China.

Red: Original text but counterintuitive and needs special attention.

"You further agree not to use the services in any manner to harass, abuse, …"

Brown: Modified text.

"Thus apple reserves right to allow its Services contents be scrutinized by … the Great Fire Wall of China"

Green: Totally irrelevant text inserted from other sources.

"Nokia said its net proceeds from the deal …"

Preprocessing: LDA Model

How to Remove common topics words from a given text?

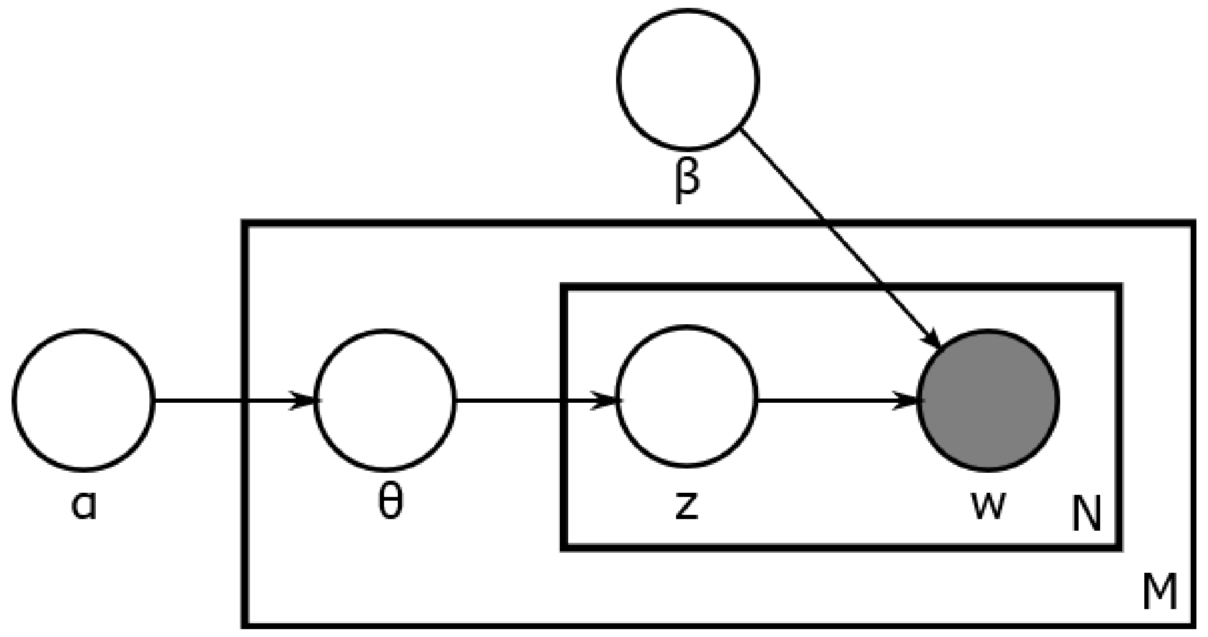
- Build a database of similar documents. Find the common topic words across the database.

Latent Dirichlet Allocation (LDA):

- A *word* **w** is the basic unit.
- A *document* is a sequence of **N** words.
- A *corpus* contains **M** documents.

Generative Process:

- Choose N ~ Poisson(ξ).
- Choose θ ~ Dir(α).
- For each N words, choose a topic z_n ~ Multinomial(θ).
- For each N words, choose a word w_n from p(w_n |z_n,β).



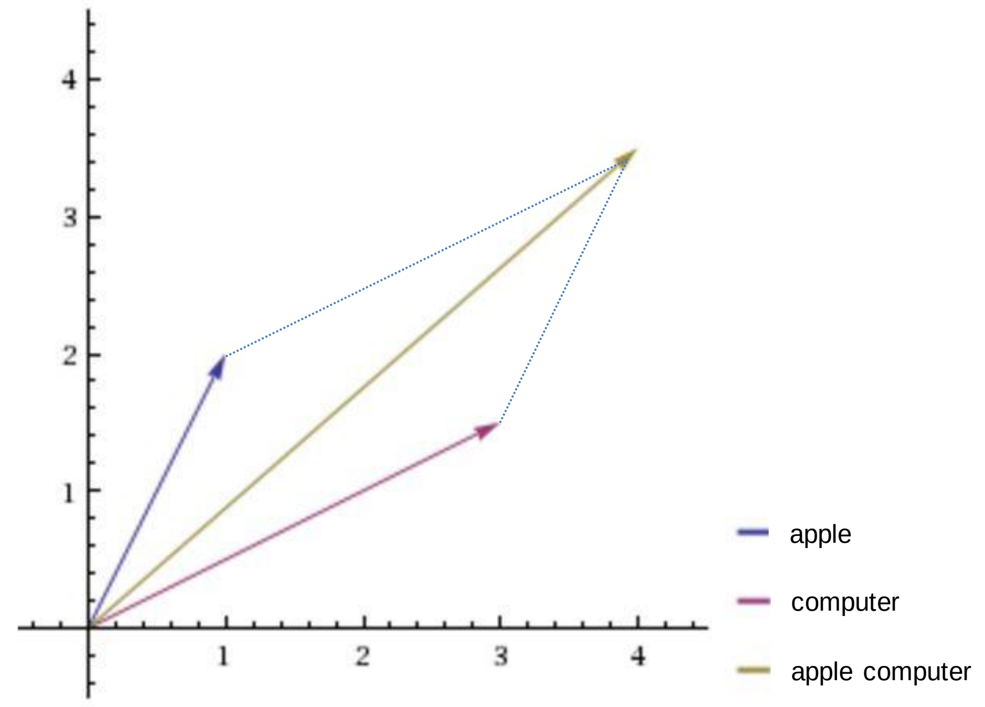
Example Topic:

Trademark + Services + May + Use + Application + Content + Will + Terms

Feature Extraction: Word2Vec

How to do arithmetic on words and sentences?

- Each word need to be represented as a vector.
- Each sentence is assumed to be sum of containing word vectors.



Example of a Word Vector:

apple = [0.68, 0.67, 0.66, …]

Example of Word Similarity:

P(apple, desktop) = 0.684

P(apple, laptop) = 0.604

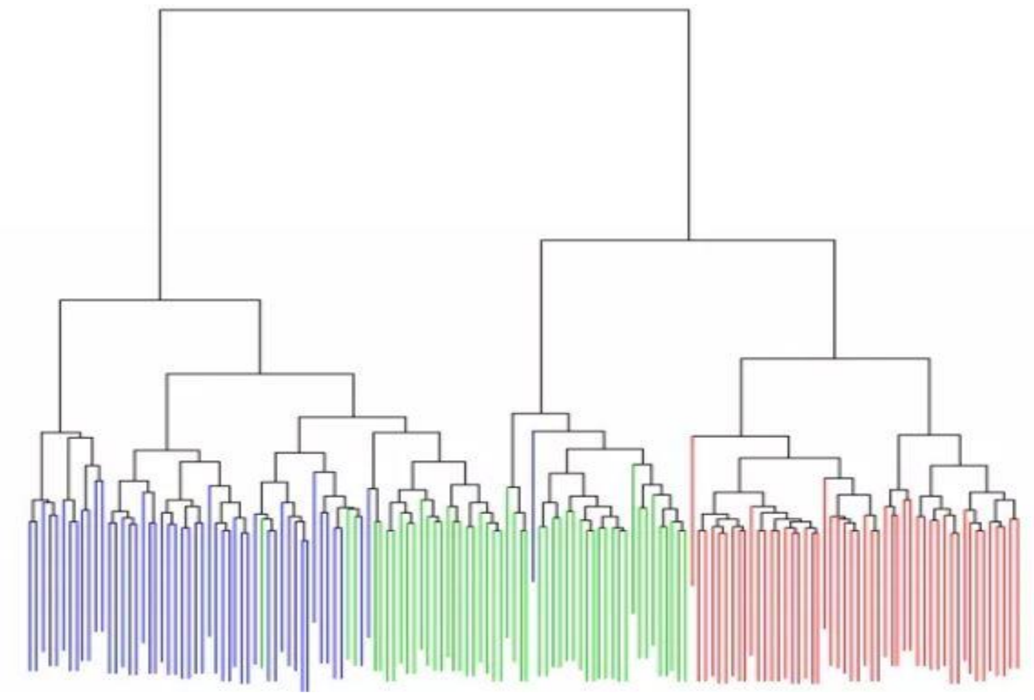
P(apple, hardware) = 0.633

…

Clustering: Agglomerative Clustering

How to find non-standard sentences that need our attention?

- **Intuition**: Cluster sentences with similar meanings; find those farthest from the centroid.
- **Challenge**: K-means gives unstable results with randomized initial centers.
- **Solution**: Agglomerative Clustering.



Green: Sentence Cluster 1

Blue: Sentence Cluster 2

Red: Sentence Cluster 3

Cluster Strategy:

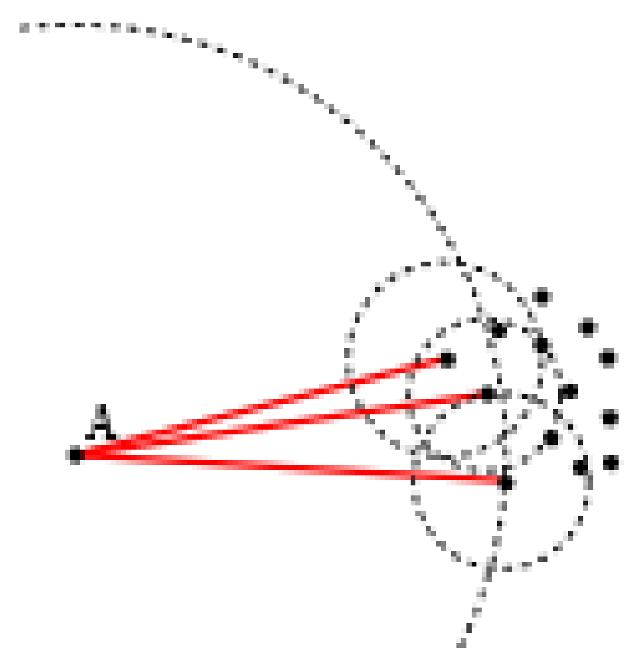
- In the beginning, every sentence is placed in a cluster.
- In each iteration, cluster with smaller average cosine distances are merged.
- At last, a sentence occupying one cluster is likely to be an anomaly.

Anomaly Detection Using LOF Model

How to find non-standard sentences that need our attention?

- **Intuition**: Sentences located with the fewest neighbors are more likely to be anomalies.
- **Solution**: Find sentences with the largest Local Outlier Factor.

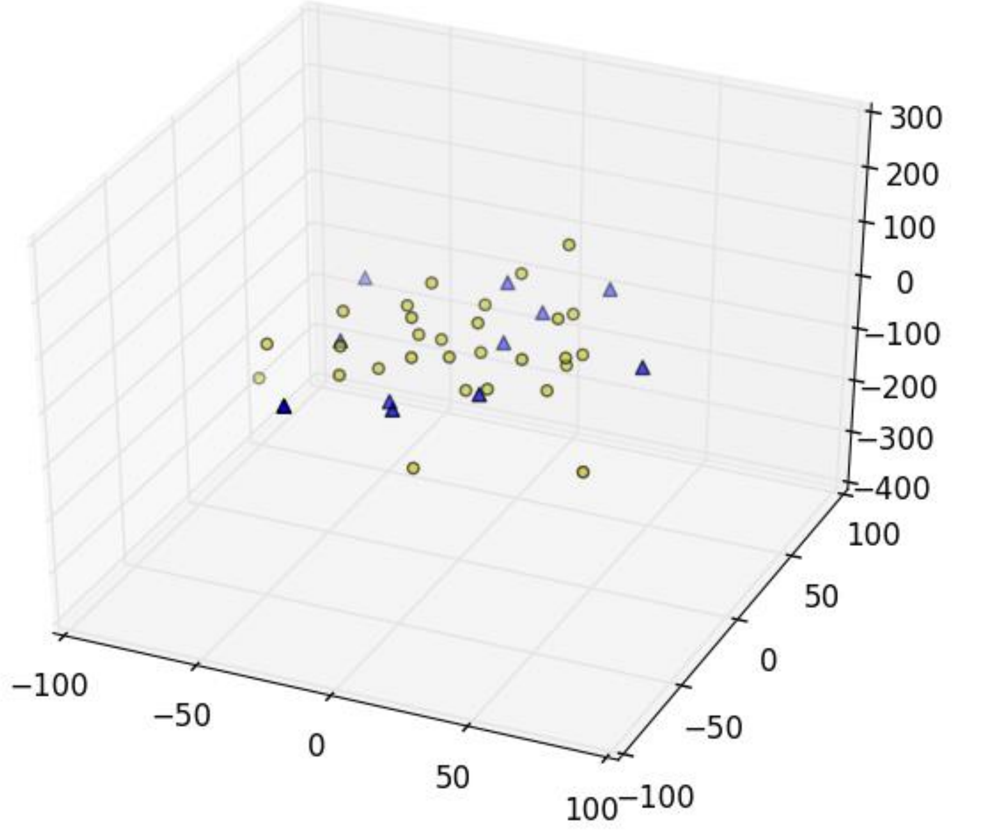
Intuition of LOF:



LOF compares the local densities of a point with its neighbors.

A has a lower density, and is more likely to be an anomaly.

Result (Plotted first 3D after t-SNE):



Blue: Anomaly sentences

Green: Normal sentences