Topic Modeling

MMA 865

Moez Ali

Smith SCHOOL OF BUSINESS | Queen's University

Smith
SCHOOL OF BUSINESS
Queen's University

TOPIC MODELING

Definitions

topic model

noun

A statistical model for discovering abstract topics.

Source: *Wikipedia*

topic

noun

The central idea of a stretch of connected discourse.

synonyms: subject, theme, issue, concern

Source: *Wikipedia*

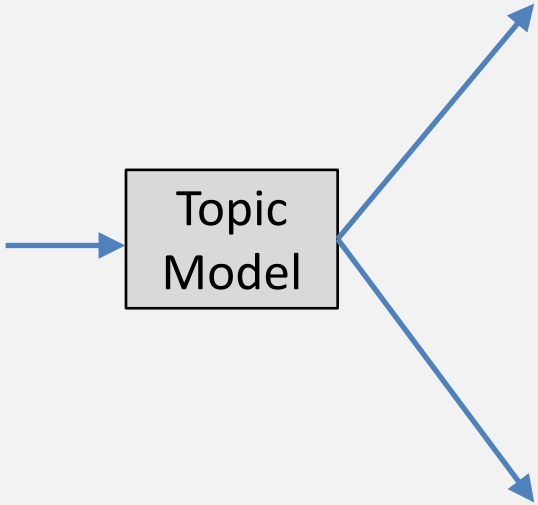
An Example

Amazon.ca's shipping business generates more cash than…

Leaders are hoping to help finance new plans to transform the LA river…

A new student loan forgiveness program would allow students to pay…

A huge transport ship that ran into the bank off the south coast of…



Topic 1	Topic 2	Topic 3
bank	river	delivery
money	bank	ship
cash	ship	transport
loan	ocean	mail
…	…	…

ID	Topic 1	Topic 2	Topic 3
1	20%	0%	80%
2	75%	25%	0%
3	100%	0%	0%
4	0%	60%	40%

What is Topic Modeling?

- Goal: automatically find the main *topics (themes)* in a set of documents
- An unsupervised ML technique
- Many algorithms: LSA (SVD), NMF, PLSA, LDA, LDA Variants
- General process:
  - User cleans/preprocesses text
  - User gives text to topic model algorithm
  - Algorithm runs for a long time
  - Algorithm discovers topics and topic assignments
  - User analyzes topics, trends, etc.
  - …
  - Profit

An Example (More Precisely)

ID	Text
1	Amazon.ca's shipping business generates more cash than…
2	Leaders are hoping to help finance new plans to transform the LA river…
3	A new student loan forgiveness program would allow students to pay…
4	A huge transport ship that ran into the bank off the south coast of…

Topic Model

Words can belong to multiple topics

Topics

	bank	money	cash	loan	river	ship	ocean	delivery	transport	mail
1	20%	15%	15%	10%	0%	0%	0%	1%	1%	0%
2	25%	0%	0%	0%	30%	20%	15%	2%	2%	0%
3	0%	0%	0%	0%	0%	35%	0%	40%	15%	10%

Topics are assigned to multiple documents

Topic Memberships

ID	Topic 1	Topic 2	Topic 3
1	20%	0%	80%
2	75%	25%	0%
3	100%	0%	0%
4	0%	60%	40%

Documents contain multiple topics

Use Case: Feature Engineering

Topic memberships become text vectorization

ID	Text
1	My dog ate my homework.
2	The cat ate my sandwich.
3	A dolphin ate the homework and the sandwich.

Topic Model

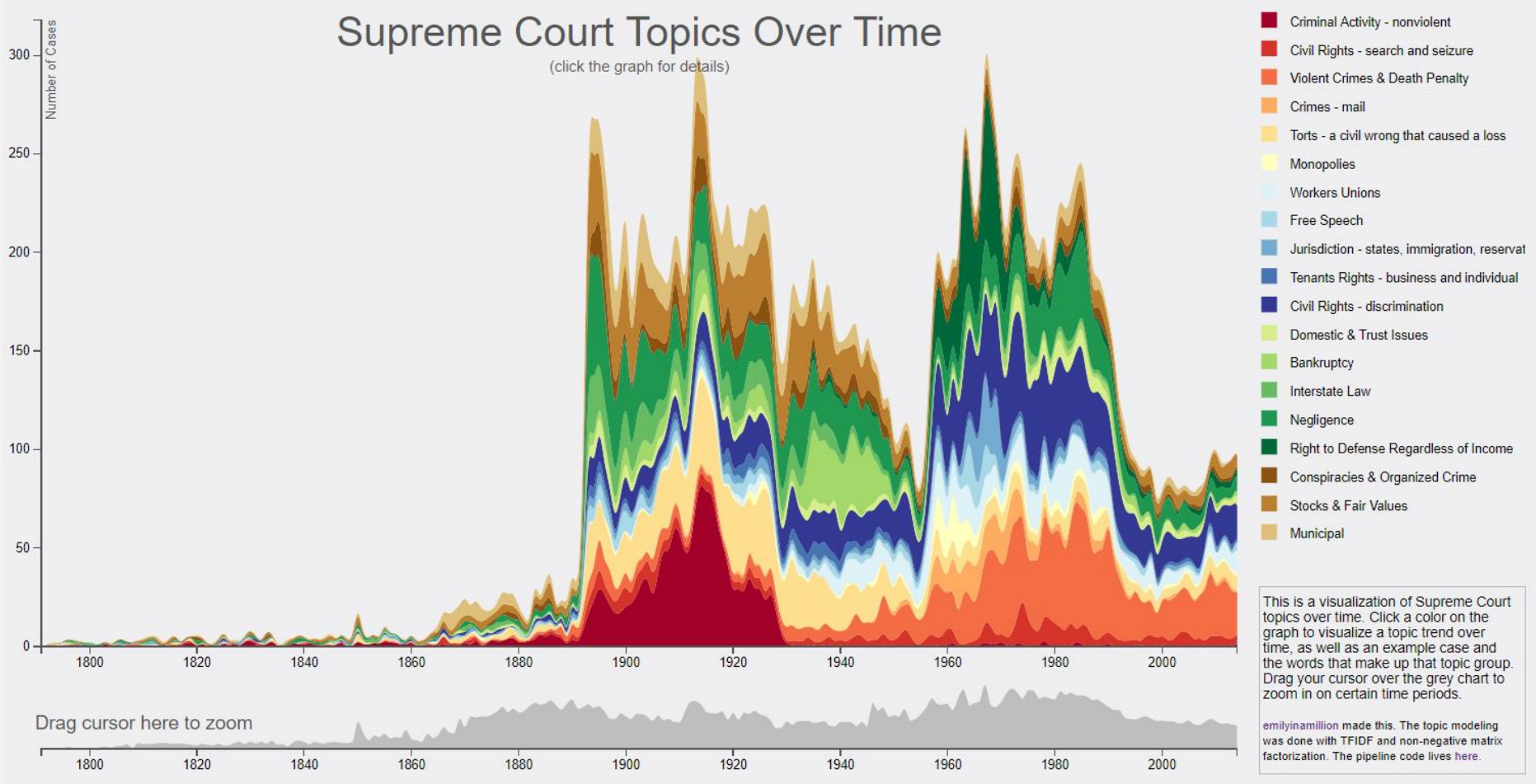
	cat	dolphin	dog	homework	sandwich
1	0.3	0.2	0.3	0	0
2	0	0	0	0.9	0
3	0	0	0	0	0.4

Topics

ID	Topic 1	Topic 2	Topic 3
1	0.5	0.5	0
2	0.5	0	0.5
3	0.4	0.3	0.3

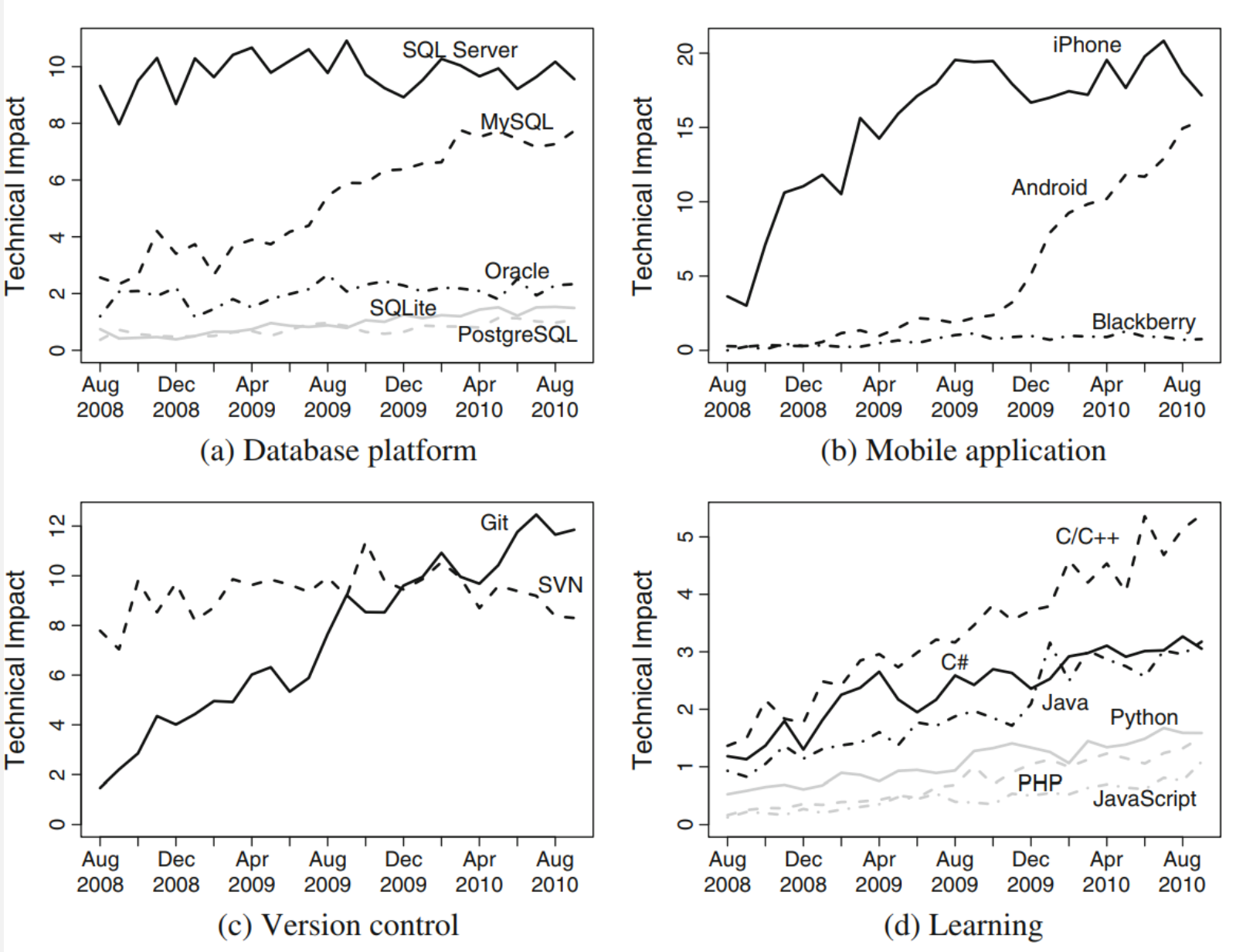
Topic Memberships

Example: Supreme Court Topics Over Time



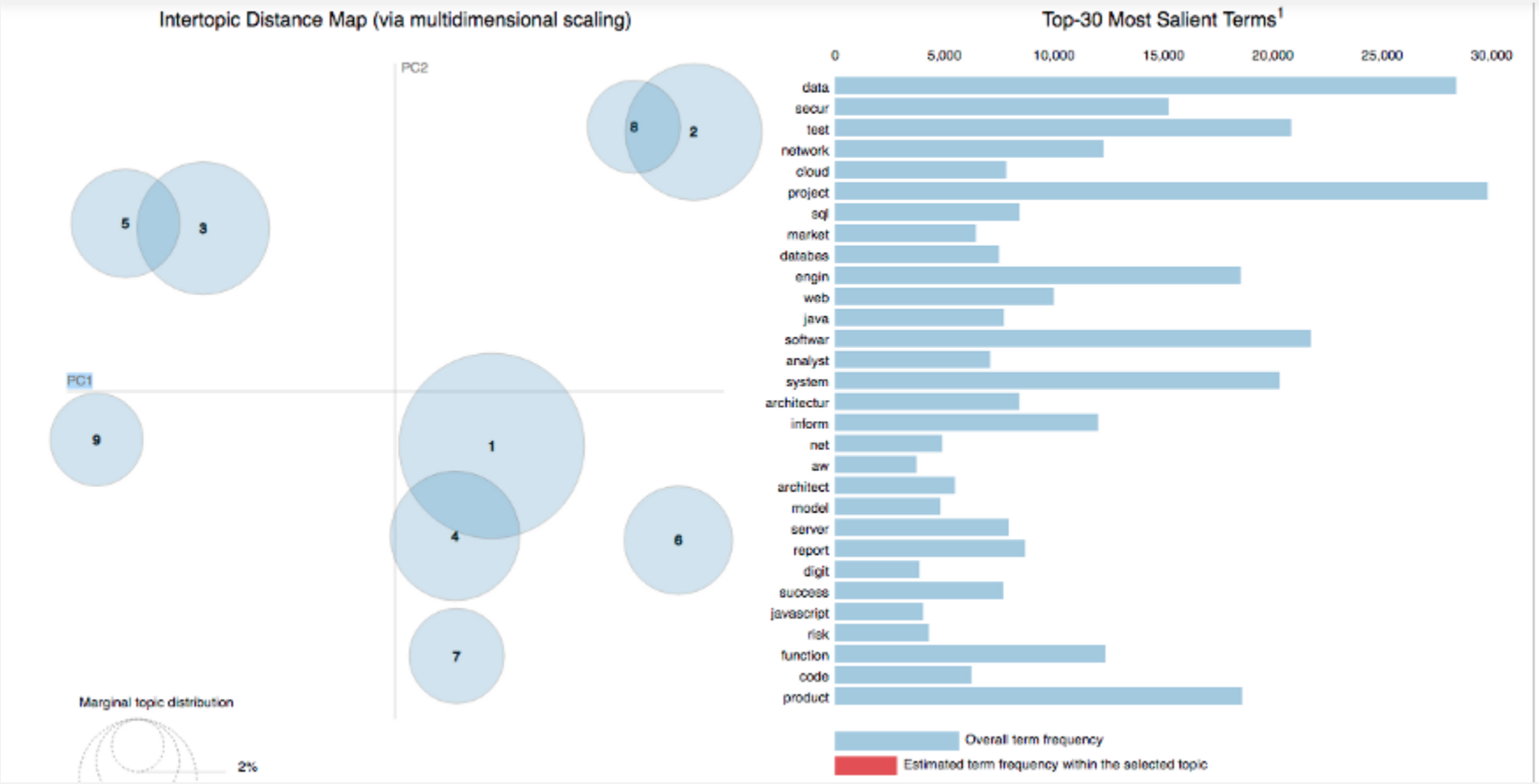
Source: emilyinamillion blog

Example: Topic Trends on Stack Overflow



Source: What are developers talking about? An analysis of topics and trends in stack overflow. Barua et al.

Visualizing Topic Models



SUMMARY