CS224C: NLP for CSS

Topic Modeling

Diyi Yang

Stanford CS

Topic Modeling

Organize the documents into a set of coherent topics

Find relationships between these topics

Understand how different documents talk about the same topic

Track the evolution of topics over time

Topic Modeling

A method of (unsupervised) discovery of latent or hidden structure in a corpus

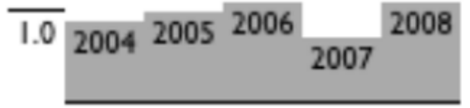
- Applied primarily to text corpora
- Provides a modeling toolbox
- Has prompted the exploration of a variety of new inference methods to accommodate large-scale datasets

problem, optimization, problems, convex, convex optimization, linear, semidefinite programming, formulation, sets, constraints, proposed, margin, maximum margin, optimization problem, linear programming, programming, procedure, method, cutting plane, solutions

Topic 54 [0.051]

decision trees, trees, tree, decision tree, decision, tree ensemble, junction tree, decision tree learners, leaf nodes, arithmetic circuits, ensembles modts, skewing, ensembles, anytime induction decision trees, trees trees, random forests, objective decision trees, tree learners, trees grove, candidate split

Topic 99 [0.066]



inference, approximate inference, exact inference, markov chain, models, approximate, gibbs sampling, variational, bayesian, variational inference, variational bayesian, approximation, sampling, methods, exact, bayesian inference, dynamic bayesian, process, mcmc, efficient

http:// www.cs.umass.edu/~mimno/icml100.html

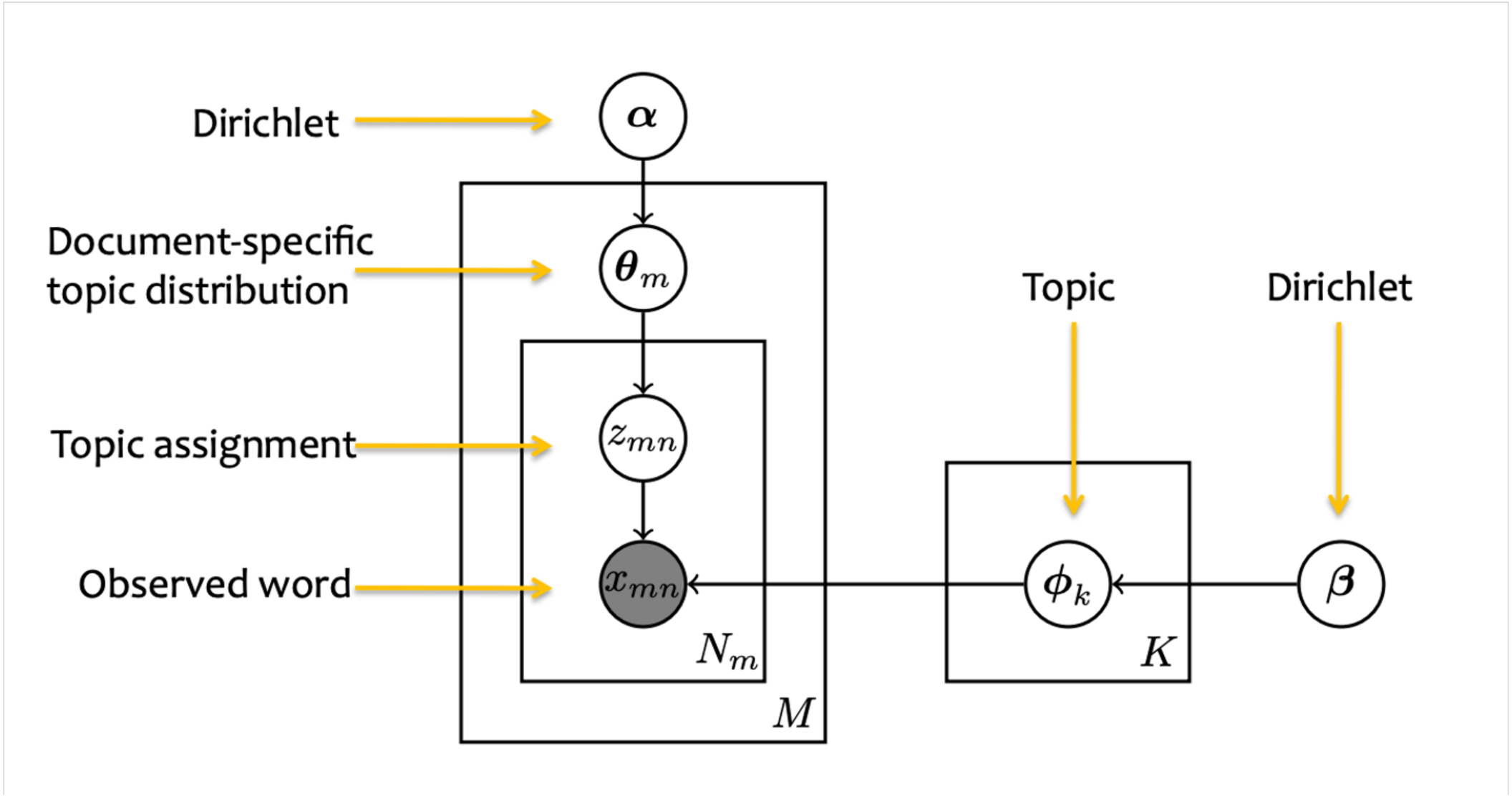
Latent Dirichlet Allocation

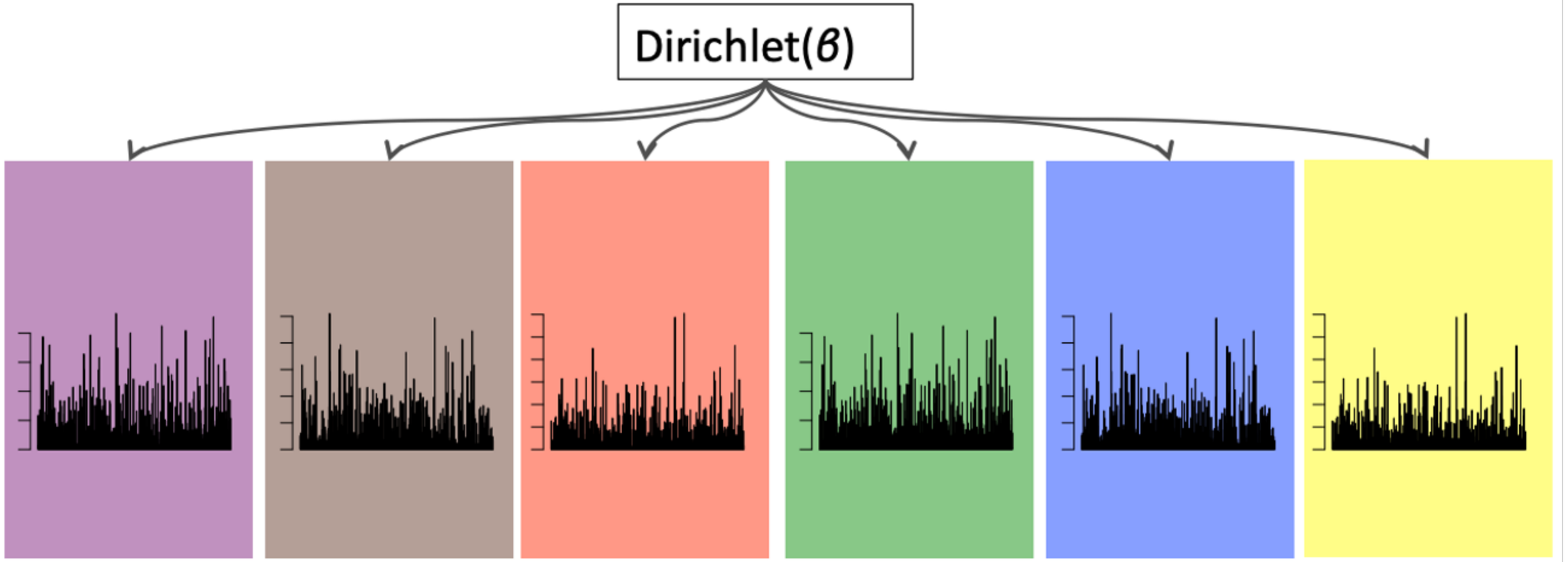
Generative Process

For each topic $k \in \{1, \ldots, K\}$:
  $\boldsymbol{\phi}_k \sim \text{Dir}(\boldsymbol{\beta})$ [*draw distribution over words*]
For each document $m \in \{1, \ldots, M\}$
  $\boldsymbol{\theta}_m \sim \text{Dir}(\boldsymbol{\alpha})$ [*draw distribution over topics*]
  For each word $n \in \{1, \ldots, N_m\}$
    $z_{mn} \sim \text{Mult}(1, \boldsymbol{\theta}_m)$ [*draw topic assignment*]
    $x_{mn} \sim \boldsymbol{\phi}_{z_{mi}}$ [*draw word*]

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." *Journal of machine Learning research* 3, no. Jan (2003): 993-1022.

Latent Dirichlet Allocation

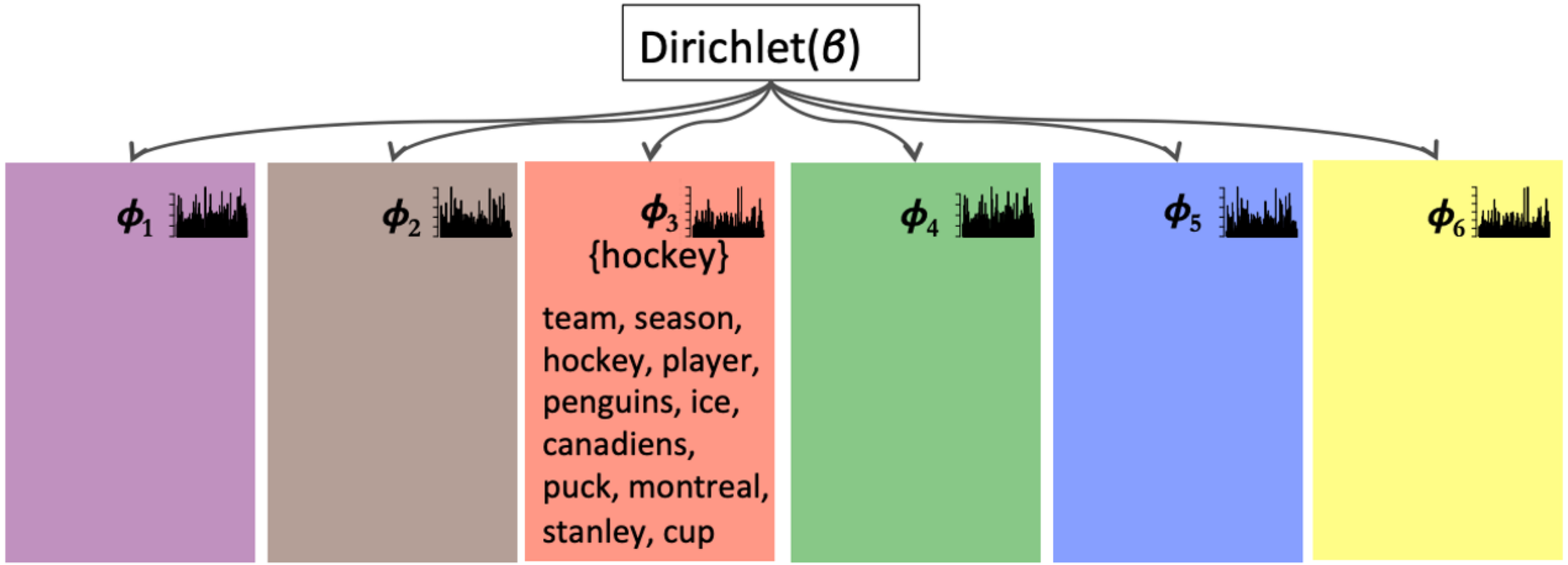




The **generative story** begins with only a **Dirichlet prior** over the topics

Each topic is defined as a **Multinomial distribution** over the vocabulary, parameterized by ϕ_k

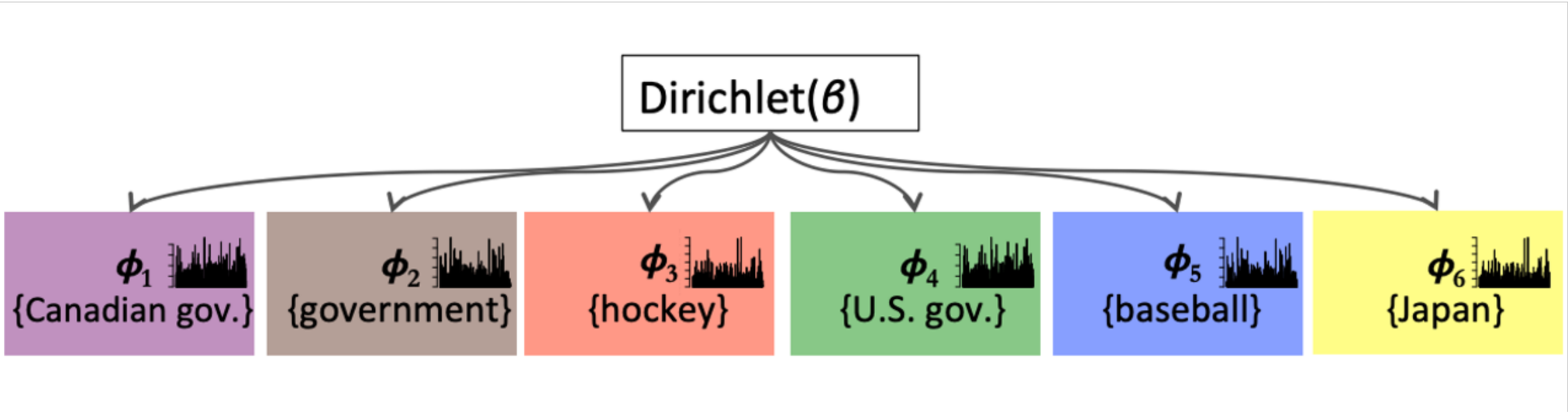
Example Credit to Matthew R. Gormley



A topic is visualized as its **high probability words.**

A pedagogical **label** is used to identify the topic.

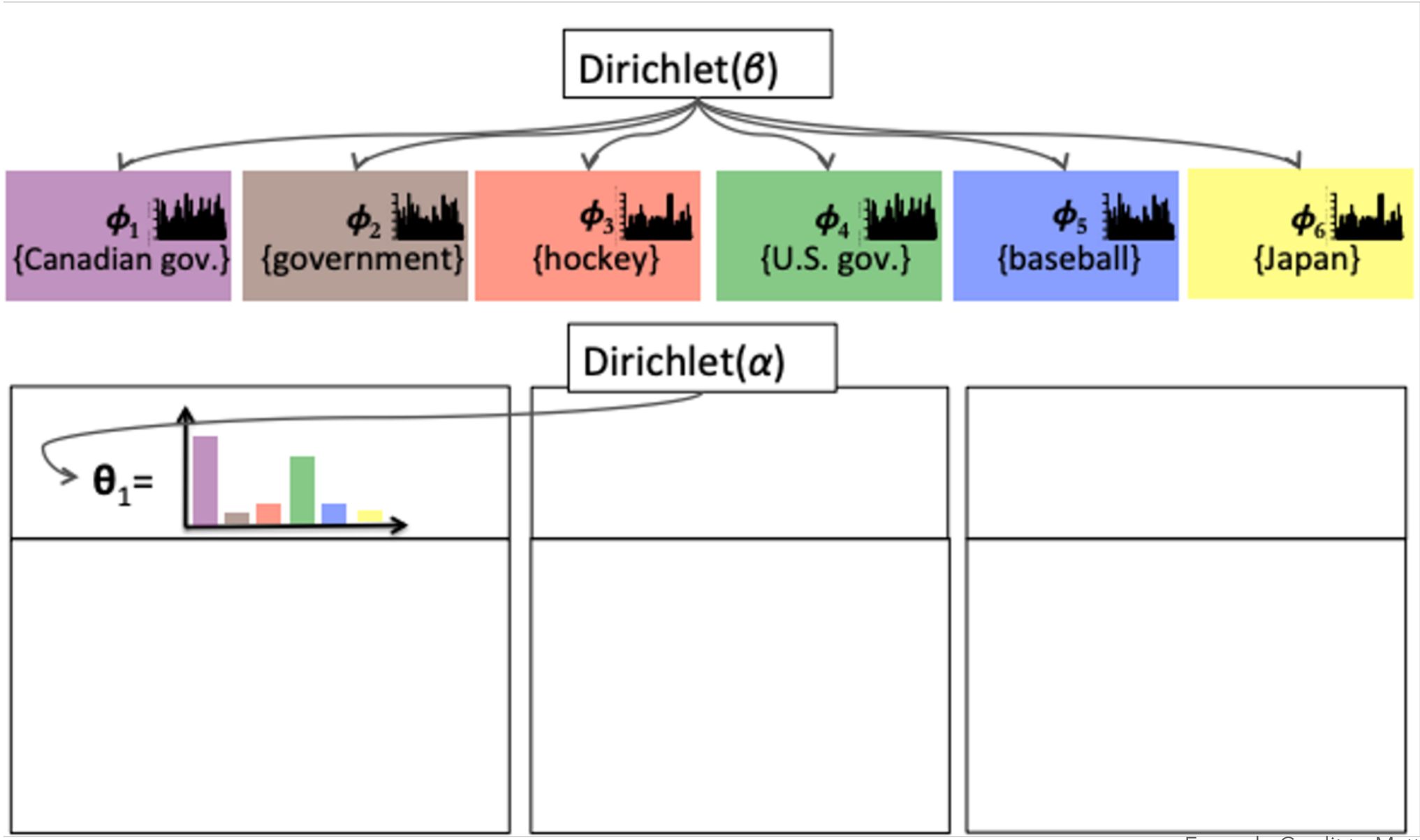
Example Credit to Matthew R. Gormley



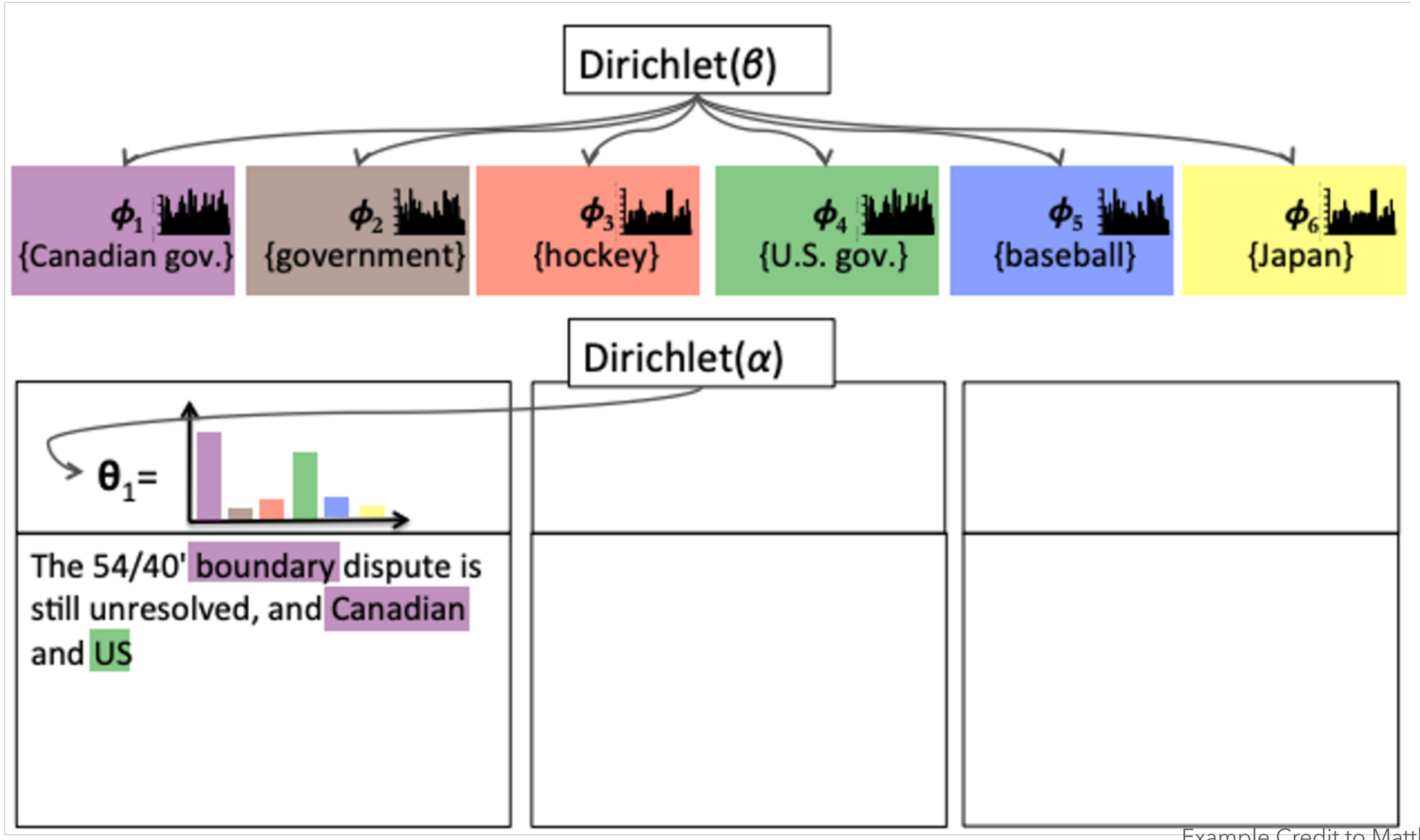
A topic is visualized as its **high probability words.**

A pedagogical **label** is used to identify the topic.

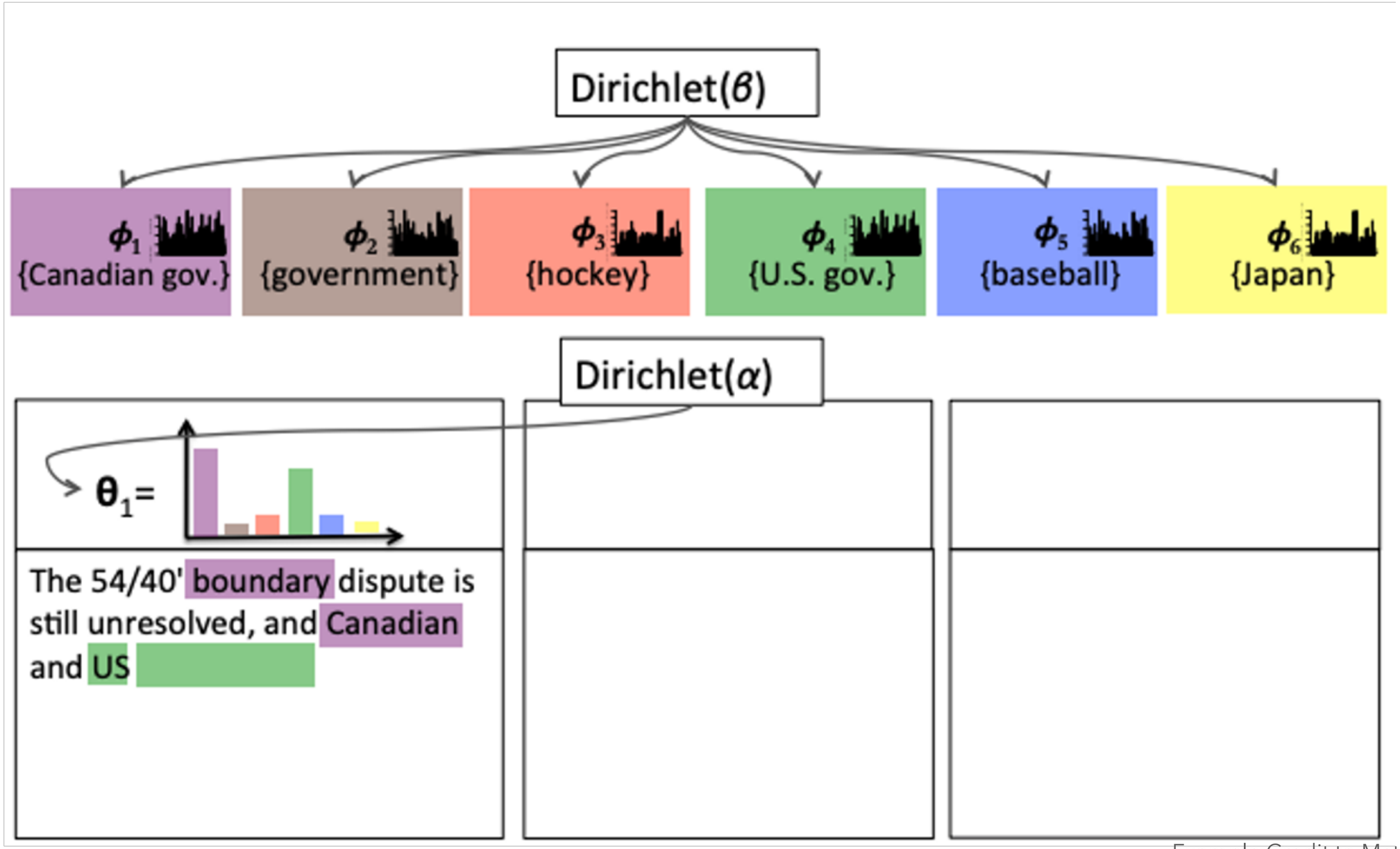
Example Credit to Matthew R. Gormley



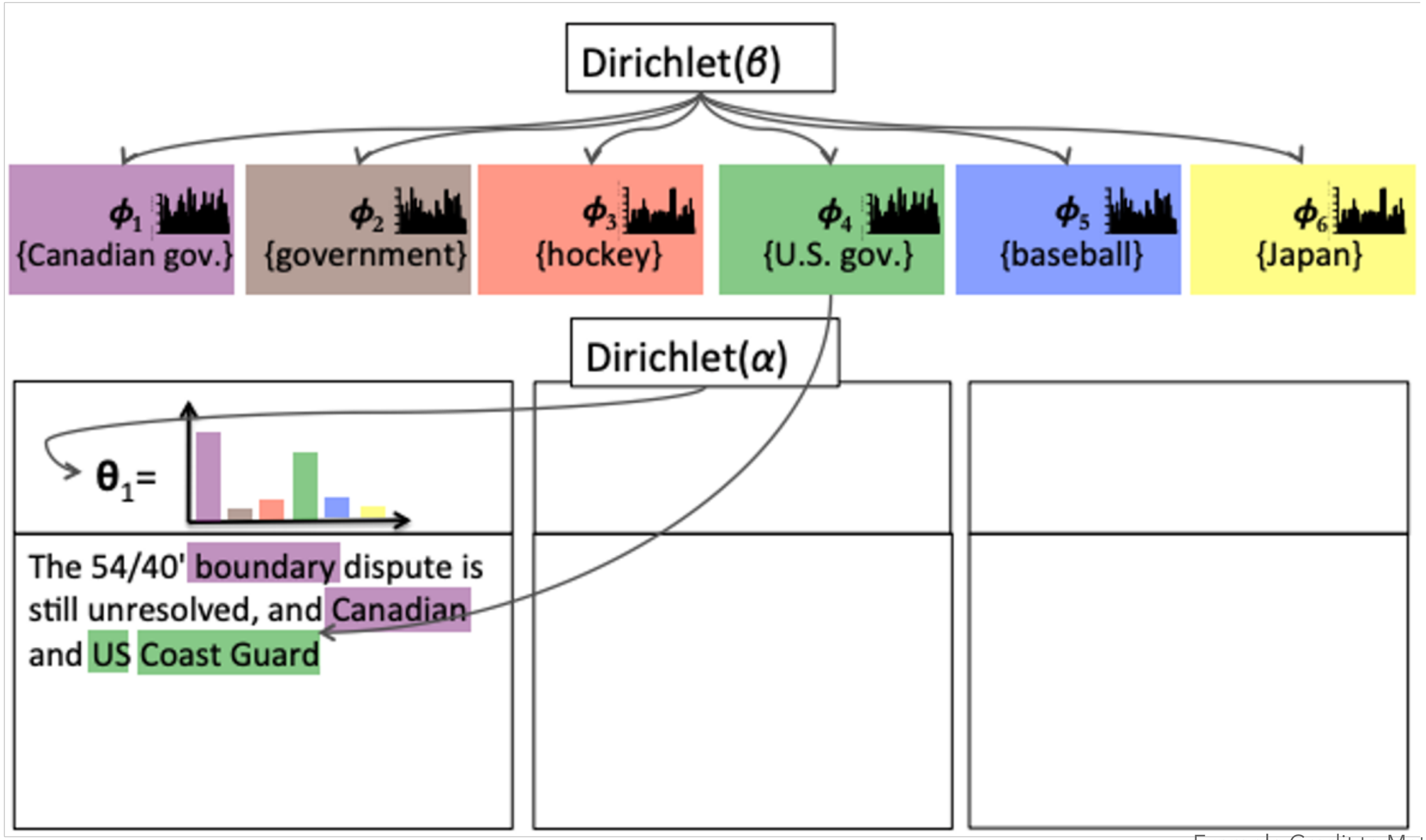
Example Credit to Matthew R. Gormley



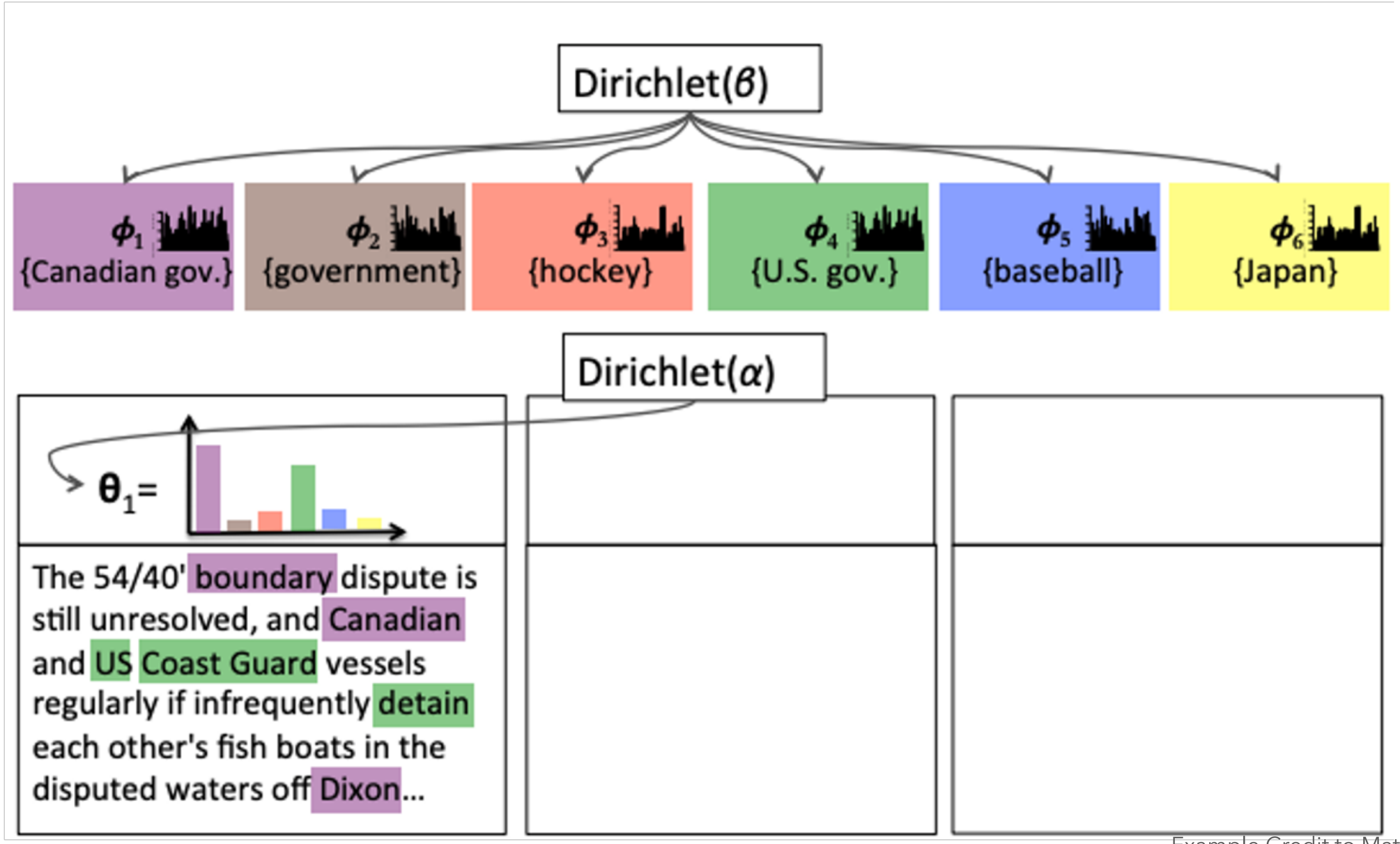
Example Credit to Matthew R. Gormley



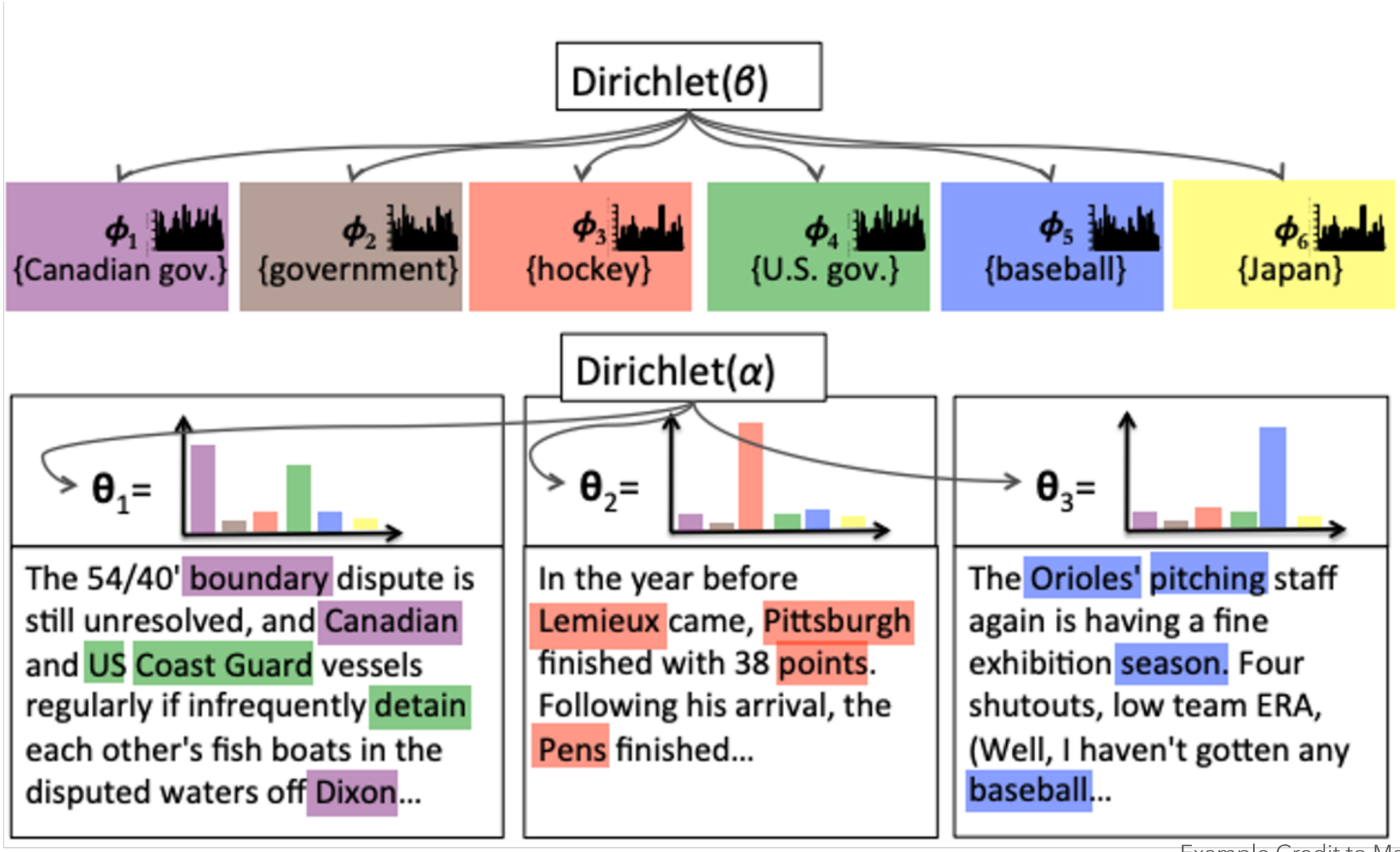
Example Credit to Matthew R. Gormley



Example Credit to Matthew R. Gormley

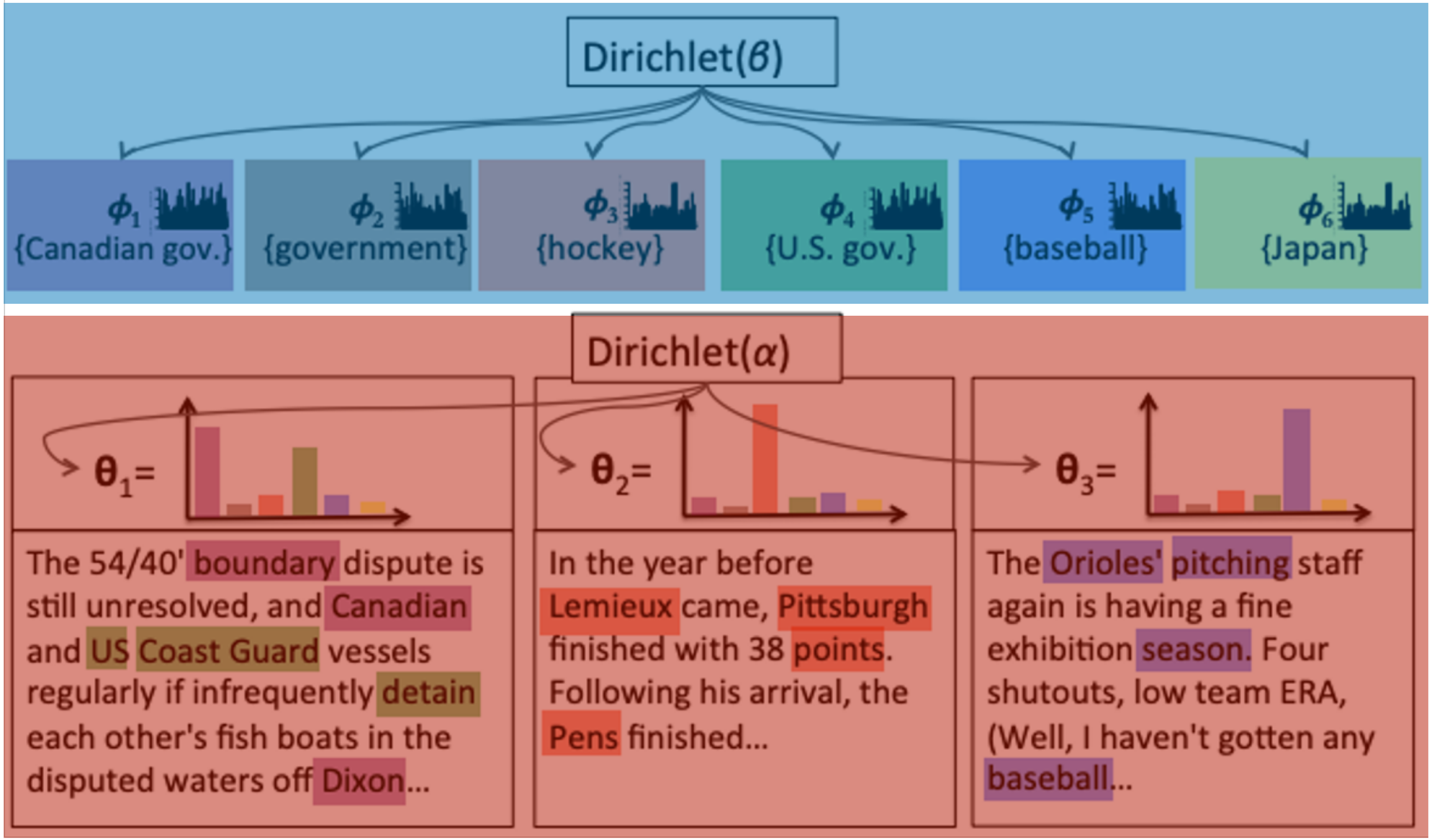


Example Credit to Matthew R. Gormley



Example Credit to Matthew R. Gormley

Distribution over words (topics)



Distribution over topics (docs)

Interpreting Topics Models

What is the meaning of each topic?

How to set the number of topics?

How to evaluate the resulting topics?

Evaluating Topic Modeling

Manual Inspection / Human judgement

- Top ranked words

Intrinsic Evaluation

- Coherence score
- Intruder test

Extrinsic Evaluation

- Downstream application

Coherence Score

Whether the words in a topic is coherent in terms of semantic similarity

UCI coherence measure $$\sum_{i<j} \log \frac{p(w_i, w_j)}{p(w_i)p(w_j)}$$

UMass coherence measure $$\sum_{i<j} \log \frac{1 + D(w_i, w_j)}{D(w_i)}$$

Mimno, David, Hanna Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. "Optimizing semantic coherence in topic models." In Proceedings of the 2011 conference on empirical methods in natural language processing, pp. 262-272. 2011.
Newman, David, Jey Han Lau, Karl Grieser, and Timothy Baldwin. "Automatic evaluation of topic coherence." In Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics, pp. 100-108. 2010.

Word Intrusion Task

Given a few randomly ordered words, find the word which is out of place or does not belong with the others, i.e., the intruder

`Dog, cat, horse, apple, pig, cow`

`Car, teacher, platypus, agile, blue, Zaire`

Chang, Jonathan, Sean Gerrish, Chong Wang, Jordan Boyd-Graber, and David Blei. "Reading tea leaves: How humans interpret topic models." Advances in neural information processing systems 22 (2009).

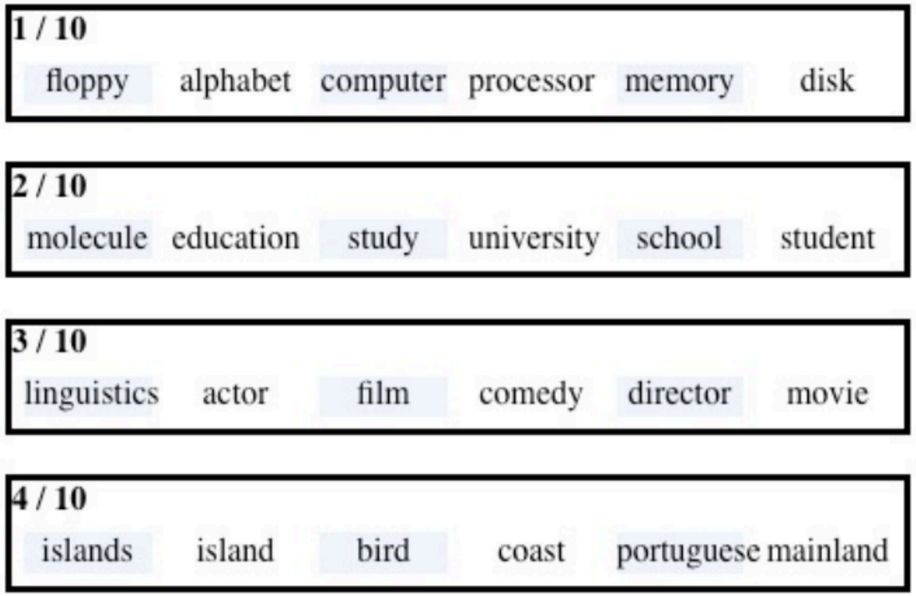
Topic Intrusion

Tests whether a topic model's decomposition of documents into a mixture of topics agrees with human judgements of the document's content

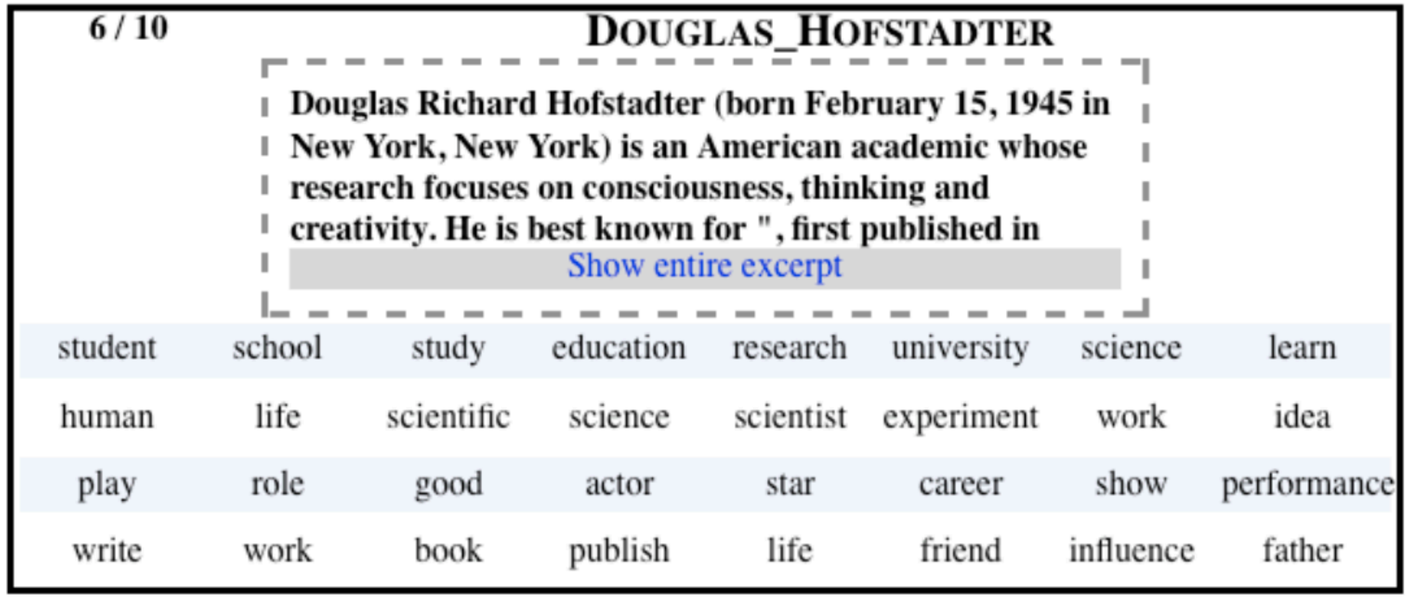
Given a title and a snippet from a document, judge which topic out of the four given topics does not belong with the document

Two Intrusion Tasks to Evaluate Topics

Word Intrusion



Topic Intrusion



What if the input text is “noisy”?

Removing non-latin characters

Filtering out stop words

e.g., “the”, “is” and “and”

Converting words to lower case?

Filtering out words with a frequency less than k

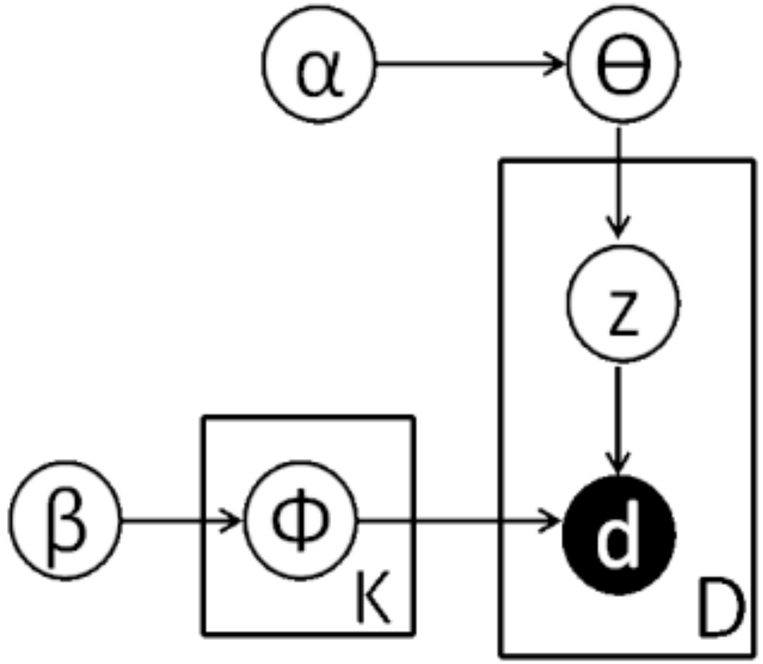
Performing stemming

…

What if the input text is short?

Dirichlet Multinomial Mixture model for short text clustering (GSDMM)

The Movie Group Process



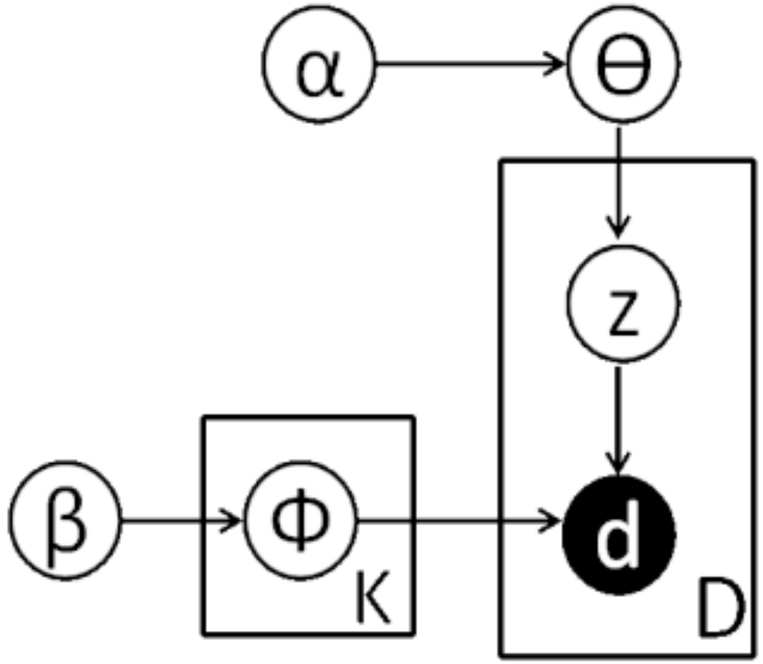
Yin, Jianhua, and Jianyong Wang. "A dirichlet multinomial mixture model-based approach for short text clustering." In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 233-242. 2014

What if the input text is short?

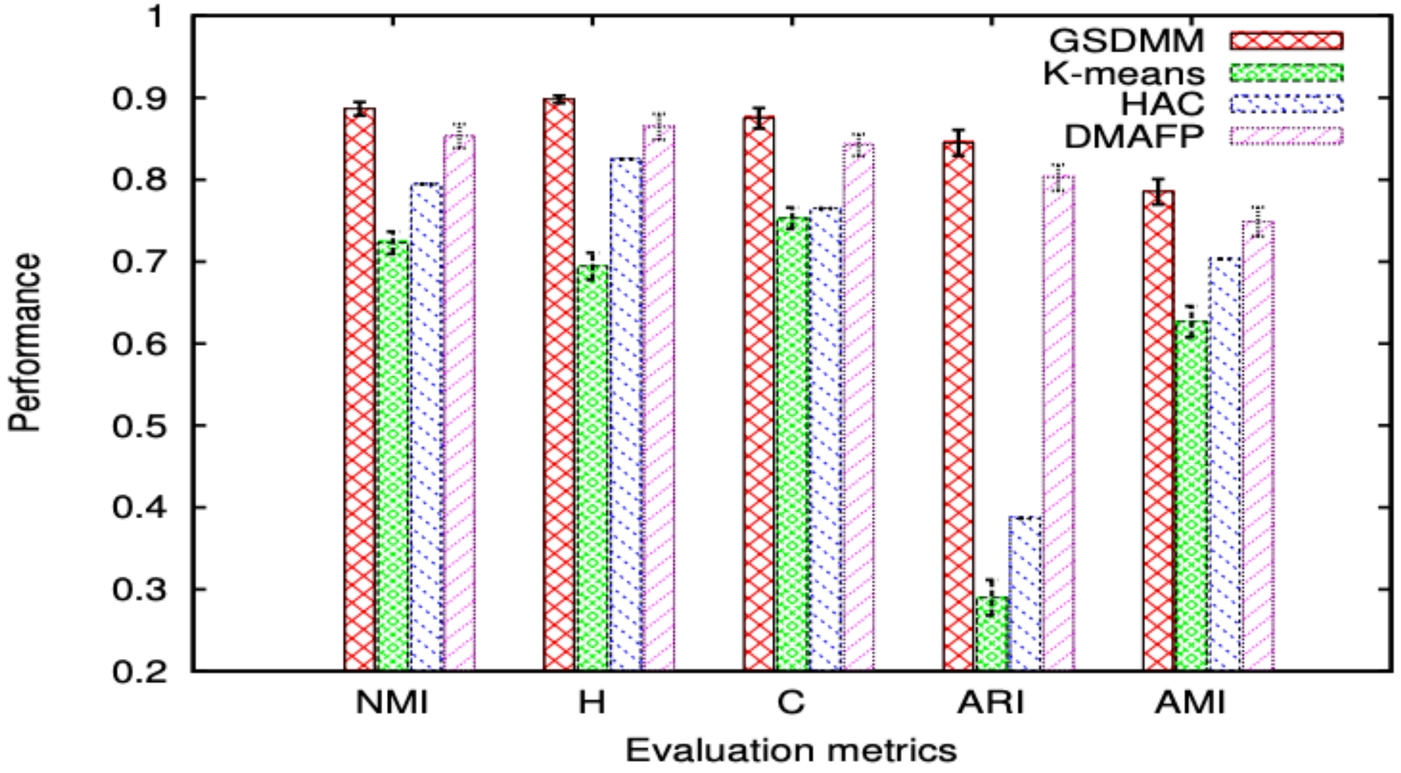
Dirichlet Multinomial Mixture model for short text clustering (GSDMM)

$$p(d) = \sum_{k=1}^{K} p(d \mid z = k) p(z = k)$$

$$p(d \mid z = k) = \Pi_{w \in d} p(w \mid z = k)$$



What if the input text is short?



Performance of the models on the TweetSet.

https://github.com/rwalk/gsdmm-rust

What if there are user priors?

"To improve topic-word distributions, we set up a model in which each topic prefers to generate words that are related to the words in a seed set"

1	company, billion, quarter, shrs, earnings
2	acquisition, procurement, merge
3	exchange, currency, trading, rate, euro
4	grain, wheat, corn, oilseed, oil
5	natural, gas, oil, fuel, products, petrol

"To improve document-topic distributions, we encourage the model to select topics based on the existence of input seed words in that document"

Jagarlamudi, Jagadeesh, Hal Daumé III, and Raghavendra Udupa. "Incorporating lexical priors into topic models." In Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, pp. 204-213. 2012.

What if there are user priors? (seededLDA)

SeededLDA allows one to specify seed words that can influence the discovered topics

topic 1: kodak, management, great, innovation, post, agree, film, understand, something, problem, businesses, changes, needs
topic 2: good, change, publishing, brand, companies, publishers, history, marketing, traditional, believe, authors
topic 3: think, work, technologies, newspaper, content, paper, model, business, disruptive, information, survive, print, media, course, assignment
topic 4: digital, kodak, company, camera, market, quality, phone, development, future, failed, high, right, old,
topic 5: amazon, books, netflix, blockbuster, stores, online, experience, products, apple, nook, strategy, video, service
topic 6: time, grading, different, class, course, major, focus, product, like, years
topic 7: companies, interesting, class, thanks, going, printing, far, wonder, article, sure

Table 2: Topics identified by LDA

topic 1: thank, professor, lectures, assignments, concept, love, thanks, learned, enjoyed, forums, subject, question, hard, time, grading, peer, lower, low
topic 2: learning, education, moocs, courses, students, online, university, classroom, teaching, coursera

Table 3: Seed words in LOGISTICS and GENERAL for DISR-TECH, WOMEN and GENE courses

topic 3a: disruptive, technology, innovation, survival, digital, disruption, survivor
topic 3b: women, civil, rights, movement, american, black, struggle, political, protests, organizations, events, historians, african, status, citizenship
topic 3c: genomics, genome, egg, living, processes, ancestors, genes, nature, epigenitics, behavior, genetic, engineering, biotechnology

Table 4: Seed words for COURSE topic for DISR-TECH, WOMEN and GENE courses

Ramesh, Arti, Dan Goldwasser, Bert Huang, Hal Daumé III, and Lise Getoor. "Understanding MOOC discussion forums using seeded LDA." In Proceedings of the ninth workshop on innovative use of NLP for building educational applications, pp. 28-33. 2014.

What if there are user priors? (seededLDA)

topic 1: time, thanks, one, low, hard, question, course, love, professor, lectures, lower, another, concept, agree, peer, point, never
topic 2: online, education, coursera, students, university, courses, classroom, moocs, teaching, video
topic 3: digital, survival, management, disruption, technology, development, market, business, innovation
topic 4: publishing, publisher, traditional, companies, money, history, brand
topic 5: companies, social, internet, work, example
topic 6: business, company, products, services, post, consumer, market, phone, changes, apple
topic 7: amazon, book, nook, readers, strategy, print, noble, barnes

Table 5: Topics identified by SeededLDA for DISR-TECH

topic 1: time, thanks, one, hard, question, course, love, professor, lectures, forums, help, essays, problem, thread, concept, subject
topic 2: online, education, coursera, students, university, courses, classroom, moocs, teaching, video, work, english, interested, everyone
topic 3: women, rights, black, civil, movement, african, struggle, social, citizenship, community, lynching, class, freedom, racial, segregation
topic 4: violence, public, people, one, justice, school,s state, vote, make, system, laws
topic 5: idea, believe, women, world, today, family, group, rights
topic 6: one, years, family, school, history, person, men, children, king, church, mother, story, young
topic 7: lynching, books, mississippi, march, media, youtube, death, google, woman, watch, mrs, south, article, film

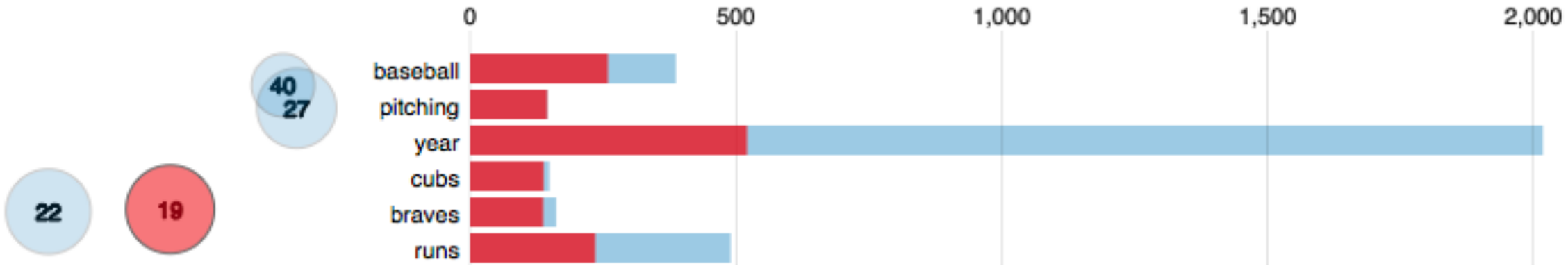
Table 6: Topics identified by SeededLDA for WOMEN

topic 1: time, thanks, one, answer, hard, question, course, love, professor, lectures, brian, lever, another, concept, agree, peer, material, interesting
topic 2: online, education, coursera, students, university, courses, classroom, moocs, teaching, video, knowledge, school
topic 3: genes, genome, nature, dna, gene, living, behavior, chromosomes, mutation, processes
topic 4: genetic, biotechnology, engineering, cancer, science, research, function, rna
topic 5: reproduce, animals, vitamin, correct, term, summary, read, steps
topic 6: food, body, cells, alleles blood, less, area, present, gmo, crops, population, stop
topic 7: something, group, dna, certain, type, early, large, cause, less, cells

Table 7: Topics identified by SeededLDA for GENE

Toolkits & Interactive topic model visualization

- Gensim
- https://github.com/bmabey/pyLDAvis
- Jupiter Notebook demo



Řehůřek, Radim, and Petr Sojka. "Software framework for topic modelling with large corpora." (2010).