Klout Topics for Modeling Interests and Expertise of Users Across Social Networks

Sarah Ellinger, Prantik Bhattacharyya, Preeti Bhargava, Nemanja Spasojevic
Lithium Technologies | Klout
San Francisco, CA
{sarah.ellinger,prantik.bhattacharyya,preeti.bhargava,nemanja.spasojevic}@lithium.com

ABSTRACT
This paper presents Klout Topics, a lightweight ontology to describe social media users’ topics of interest and expertise. Klout Topics is designed to: be human-readable and consumer-friendly; cover multiple domains of knowledge in depth; and promote data extensibility via knowledge base entities. We discuss why this ontology is well-suited for text labeling and interest modeling applications, and how it compares to available alternatives. We show its coverage against common social media interest sets, and examples of how it is used to model the interests of over 780M social media users on Klout.com. Finally, we open the ontology for external use.

KEYWORDS
taxonomy; topic ontology; topic modeling; social media; user interest models, user expertise models

2 MOTIVATION
How can the choice of a candidate topic set impact the success of an interest modeling application? Consider the topics that could be assigned to the Twitter profile of machine learning expert Andrew Ng. For an interest modeling application to assign meaningful labels to this profile, it would need a topic set containing concepts like “AI”, “Coursera”, and “Stanford”. But is there value in representing “AI” and “Artificial Intelligence” as distinct topics? Should the topic set contain entities like “Baidu AI Group”, and would we expect them to be successfully modeled? Should the topic set contain abstract concepts like “Thought” or “Power”, and would those convey any meaning when assigned? If the topic set included negatively-viewed concepts such as “Terrorist Propaganda” or “Spam”, and those topics were assigned based on a post like Figure 1, would we risk implying that the user is either a terrorist or a spammer? These risks increase when modeling interests over large numbers of social media profiles, and make it advisable to use a carefully curated topic set.

3 BEST PRACTICES
Through the practice of modeling social media interest and expertise topics on Klout.com, we have identified the following best practices for an interest modeling topic set:

(1) **Contain concepts that can be meaningful interests of a person or topics of a document.** One can reasonably say “This social media user has in-depth knowledge of geometry” or “this article is about geometry”, which

---

1 http://business.twitter.com/en/targeting/interest.html
2 https://support.google.com/ads/answer/2842480?hl=en
3 http://twitter.com/AndrewYNg
imply that “Geometry” is a good candidate topic. However, one cannot reasonably say “This social media user has in-depth knowledge of rectangles”, which tells us that assigning “Rectangles” as a topic will likely be seen as a failure of the modeling application.

(2) Be limited to topics, rather than text formats, user demographics, or other attributes. We are supporting the task of determining what a given piece of text is about and then aggregating that topic knowledge to label a social profile; it is out of our scope to determine non-topical attributes of the user or the text.

(3) Contain sufficient depth and range of concepts to cover the wide variety of topics discussed on social media. A greater number of distinct, meaningful topics allows for more meaningful targeting or labeling of users. Narrower subcategories such “Gospel Music” or “Affordable Care Act” can be more useful to an application than higher-level categories like “Music” or “Politics”.

(4) Be small enough for ease of maintenance. While a too-small topic set has limited utility, as discussed above, a very large topic set is challenging to curate. In addition, a very large topic set can be expensive to model over large amounts of social media content.

(5) Be up-to-date with topics likely to be discussed on social media. Certain domains, such as technology, politics and popular culture, evolve rapidly; a topic set from even a year ago is likely to have problematic gaps.

(6) Be readable by end users. If the modeled topics are to be exposed in an application, they should be named with clarity and brevity. (e.g., “Humidifiers” is preferable to “Humidifier (Product)”).

(7) Be structured hierarchically to leverage connections between concepts. This allows the modeling application to infer knowledge upward in the graph (e.g., to understand that a user who is interested in “Chicago Cubs” will also be interested in “Baseball”).

(8) Support multiple languages. A single-language topic set will limit the final application.

(9) Include knowledge base entity ids, for data extensibility. Mapping topics to entities in a knowledge base such as Wikipedia or Freebase is useful for a number of NLP tasks, such as entity disambiguation.

4 ALTERNATIVE TOPIC SETS

We find that the alternative sources of topic labels fall into the following categories:

(1) Formal ontologies developed for the Semantic Web such as OpenCyc\(^4\), OntoLingua\[^3\], Schema.org\[^5\], etc. These are larger and more complex than the social media topic domain requires, and would require extra processing to apply. For example, the OpenCyc node “Game” is, among other things, an instance of the “type of temporally stuff-like thing”, which is a complex category out of the scope of topic modeling tasks.

(2) Be structured hierarchically to leverage connections between concepts. This allows the modeling application to infer knowledge upward in the graph (e.g., to understand that a user who is interested in “Chicago Cubs” will also be interested in “Baseball”).

(3) Support multiple languages. A single-language topic set will limit the final application.

| Root                | Nodes | Edges |
|---------------------|-------|-------|
| Arts and Humanities | 565   | 625   |
| Food and Drink      | 265   | 113   |
| Government and Politics | 484 | 527   |
| Health and Wellness | 374   | 422   |
| News and Media      | 157   | 164   |
| Science and Nature  | 635   | 623   |
| Sports and Recreation | 827  | 894   |
| Location            | 912   | 1019  |
| Fashion             | 242   | 259   |
| Education           | 648   | 656   |
| Hobbies             | 92    | 110   |
| Entertainment       | 5779  | 5325  |
| Business            | 869   | 884   |
| Technology          | 1131  | 1196  |
| Lifestyle           | 635   | 681   |
| Travel and Tourism  | 73    | 79    |

Table 1: Root categories of KT, with counts of nodes and edges

5 KLOUT TOPICS CHARACTERISTICS

We present Klout Topics, a lightweight curated ontology designed to model user and text topics across social networks. The latest version can be accessed on GitHub\[^10\].

5.1 Methodology

KT was initially bootstrapped from a set of 140K Freebase topic nodes, selected by matching popular keywords from social media text. That set was curated down to 20K candidates by removing nodes that were not sufficiently popular on social media (“Australian Desert Raisin”) or not sufficiently

\[^4\] http://opencyc.org/ \[^5\] http://schema.org/

\[^6\] https://www.wikidata.org/ \[^7\] http://wiki.dbpedia.org/

\[^8\] http://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/

\[^9\] https://www.dmoz.org/ \[^10\] https://github.com/klout/opendata/tree/master/klout
intuitive for application use (“Comments”). Relationships within those 20K nodes were then inferred from Freebase and verified by curators. With ongoing curation, KT reached its current size of 8K nodes and 13K edges, with newly relevant topics added and obsolete ones removed.

5.2 Scope
As users on social networks interact with a variety of topics and domains, KT aims to include any concept that:

(1) is specific enough to meaningfully describe a user’s interest or the primary topic of a piece of text (e.g., “Life” or “Growing” are too broad, but “Biology” or “Child Development” should be included).

(2) is general enough to be shared by a community of interest on social media, with content that can be detected and recommended. A topic that is theoretically possible, but lacks users or content (e.g. “Scandinavian Death Metal Played on Kazoos”) will not be included.

(3) is not redundant with topics already present (“Child Development Milestones” does not need to be a separate topic from “Child Development”).

(4) is not illegal or offensive to a general audience.

5.3 Components and definitions
5.3.1 Topic nodes. Each topic node in KT contains:

(1) topic_id: a unique numerical identifier for the topic.

(2) slug: a human-readable English label, usable in a URL.

(3) metadata: includes both the Wikidata entity ID and Freebase machine_id of the closest corresponding entity. This facilitates NLP tasks such as entity disambiguation and linking [1].

5.3.2 Topic Edges. KT is structured as a directed graph with all branches connected to one or more of the 16 root categories in Table 1. Each edge indicates a parent-child relationship where a child may be a subclass of the parent (e.g., “Philosophy” > “Ethics”), or a notable entity within the topic context (“Philosophy” > “Friedrich Nietzsche”). A child may have multiple parents in different branches of the graph, and parents may be at different distances from the root.

Each edge contains:

(1) edge_id: a unique numerical identifier for the edge.

(2) source_topic_id: the topic_id of the parent topic.

(3) destination_topic_id: the topic_id of the child topic.

5.3.3 Topic Display Names. A topic node is associated with the following fields to define how it should be displayed:

(1) topic_id: the unique identifier for the topic.

(2) language: the language of the given display_name.

(3) display_type: a flag to indicate whether the topic should be displayed in the given language.

(4) display_name: the most correct and human-readable form of the topic name, in the given language.

In order for KT to be used in consumer-facing applications, its editorial voice strives to be transparent (avoiding neologisms and domain-specific jargon), neutral or positive in tone (avoiding derogatory speech and using self-identified terms for groups of people).

### Table 2: Frequently and infrequently assigned Klout Topics.

| Frequent Interest | Rare Interest | Frequent Expertise | Rare Expertise |
|-------------------|---------------|-------------------|---------------|
| Twitter | Andorra | Spare | Aplia |
| Music | AstraZeneca | Actors | AstraZeneca |
| Actions | Queen's City | Celebrities | Big Band |
| Software | Inbound Marketing | Media | Aplia |
| Television | Category Theory | Spare | Biochemistry |
| Internet | Youtube | Tel Aviv | Tel Aviv |
| Sports | Marketing Automation | Soccer | Inbound Marketing |
| Media | Bioethics | Inheritance | Category Theory |
| Celebrities | Objective-J | Singing | Singing |
| YouTube | Skiers | Internet | Objective-J |
| Entertainment | Osbourne Impala | Television | Skiers |
| Film | Assisted Living | Human | Dreamworks |
| History | Bavaria | Politics | Marketing Automation |
| Politics | Tapestry | Portugal | Netflix Comedy |

### Table 3: KT size and depth vs. Twitter and Google interests.

| Interest | KT | Twitter | Google |
|----------|----|---------|--------|
| Top-level | 16 | 25 | 11 |
| Nodes | 8028 | 350 | 2042 |
| Depth | 6 | 2 | 9 |

### Table 4: Frequent Klout Topics not in Twitter or Google sets.

| Interest | Expertise |
|----------|-----------|
| Twitter | Twitter |
| YouTube | YouTube |
| Feminism | Singing |
| Islam | Islam |
| Donald Trump | Feminism |
| Islam | Mecca |
| Leadership | Donald Trump |
| MTV | Riyadh |
| Inspiration | MTV |
| Facebook | Justin Bieber |
| One Direction | Inspiration |
| Instagram | Instagram |
| Nature | K-Pop |
| Snapchat | CNN |
| LGBT | LGBT |

5.4 Internationalization
A core assumption of KT is that the concept encoded in a node is not language- or region-specific, but can be shared across languages. Rather than maintaining separate topic sets for each language, therefore, a single multi-lingual instance can be used. This approach allows for direct comparison of the interests of users in different languages. For flexibility in the application, however, we include a flag to indicate whether a topic should be displayed in a given language.

Currently, the following languages are supported: Arabic (KSA), English (USA), French, German, and Spanish (EU).

6 KLOUT TOPICS COVERAGE VERSUS ALTERNATIVES

Here we discuss how the coverage of Klout Topics compares to other topic sets used for modeling user interests. Table 3 compares the topical coverage of KT versus Twitter Interests[7] and Google ad categories[2]. While the number of top-level categories is comparable across all three, KT contains the most overall nodes, as well as a significantly deeper structure than Twitter Interests and a comparable one to Google Ads.

In practice, Klout Topics also allows us to model both widely-shared and relatively narrow interests. Table 2 shows a sample of Klout Topics assigned to 50M or more profiles.
Table 5: Sample Twitter and Google interests not in KT.

| Twitter                                      | Google                                        |
|----------------------------------------------|-----------------------------------------------|
| Dresses and skirts                           | ATMs and Branch Locations                     |
| Empty nesters                                | Bottled Water                                 |
| Kids’ apparel                                | Clip Art and Animated GIFs                    |
| Men’s beachwear                              | Men’s Interests                               |
| Venues                                       | Offbeat                                       |
| Women’s accessories                          | Puzzles and Brainteasers                      |
| Women’s beachwear                            | Song Lyrics and Tabs                          |
| Women’s jeans                                | Special Occasions                             |
| Women’s pants                                | Vacation Offers                               |
| Women’s tops                                 | World Localities                              |

out of over 780M scored on Klout.com (“Frequent Interest/Expertise”), as well as a sample of Klout Topics assigned to 10K or fewer (“Rare Interest/Expertise”), as of June 2017.

Table 4 shows popular Klout Topics (assigned to 10M or more profiles as of late June 2017) that are not present in either the Twitter or Google sets. Other examples of topics unique to KT include some brands (“Converse”, “Pizza Hut”), news-driven concepts (“Affordable Care Act”, “Criminal Justice Reform”), and narrow subconcepts (“West Papua” in addition to “Indonesia”, “Aquariums” in addition to “Zoos”, “DevOps” in addition to “Programming”). Conversely, Table 5 shows a sample of Twitter and Google topics not present in KT. Analysis finds 24 Twitter and 154 Google topics not represented in some form in KT. Many of these missing topics are product subtypes (“Women’s Beachwear”) or categories of URLs (“Skins Themes and Wallpapers”).

7 CASE STUDIES

Here, we show examples of how KT is used to model expertise topics of over 780M social media users on Klout.com, using the Klout topic assignment system ([6], [5]). These topic scores are also available via the Twitter PowerTrack API11.

Figure [2a] shows the expertise topics displayed on the Klout.com profile of Andrew Ng, which range from the broad (“Computer Science”) to the relatively narrow (“Apache Spark”). Figure [2b] shows the Klout.com profile of Senator Kamala Harris, and includes many topics specific to her state of California as well as some relating to government and politics. Figure [2c] shows the Klout.com profile of one of the authors of this paper, and includes a mix of modeled expertise topics (“Eventbrite”, “Klout”) as well as non-modeled topics from KT that the user has selected (“Parenting”).

8 CONCLUSION AND FUTURE WORK

In this paper, we have described the importance of a robust topic set for text labeling and topic modeling applications taking input from social media; the topic set should be lightweight, human-readable, span multiple domains of knowledge, and be mapped to a commonly used knowledge base for extensibility to natural language processing tasks. We then discussed the drawbacks of some of the currently available options for this purpose, and proposed Klout Topics (KT) instead. We gave an overview of KT’s purpose, characteristics, and coverage; and finally, we gave examples of how it can be used to model a range of topics based on a person’s social network activity.

REFERENCES

[1] Preeti Bhargava, Nemanja Spasojevic, and Guoning Hu. 2017. Lithium NLP: A System for Rich Information Extraction from Noisy User-generated Text on Social Media. In Proc. of the 3rd Workshop on Noisy User-generated Text, 131 – 139.
[2] Google. 2017. Google Ad Topics. https://support.google.com/ads/answer/2842480?hl=en. (2017).
[3] Thomas R Gruber. 1992. Ontolingua: A mechanism to support portable ontologies. Vol. 27. Citeseer.
[4] Martin Hepp. 2006. Products and Services Ontologies: A Methodology for Deriving OWL Ontologies from Industrial Categorization Standards. International Journal on Semantic Web & Information Systems 2, 1 (2006).
[5] Nemanja Spasojevic, Prantik Bhattacharyya, and Aditya Rao. 2016. Mining half a billion topical experts across multiple social networks. Social Network Analysis and Mining 6, 1 (2016), 63.
[6] Nemanja Spasojevic, Jinyun Yan, Aditya Rao, and Prantik Bhattacharyya. 2014. LASTA: Large Scale Topic Assignment on Multiple Social Networks. In Proc. of ACM Conference on Knowledge Discovery and Data Mining (KDD) (KDD ’14).
[7] Twitter. 2017. Ad targeting best practices for Twitter. http://business.twitter.com/en/targeting/interest.html. (2017).

11 http://support.gnip.com/apis/powertrack/