Enriching iTunes App Store Categories via Topic Modeling

Svitlana Vakulenko, Oliver Müller, Jan vom Brocke
MODUL University Vienna, University of Liechtenstein

UNIVERSITÄT LIECHTENSTEIN

Motivation

The rapid growth of mobile app markets creates challenges for consumers, developers, and marketplace providers. As the number of apps offered is constantly rising, it becomes more and more difficult for all the stakeholders to gain an overview of the app market.

Stakeholders and their goals:

- **End-users**: easily discover apps they need and like to aquire.
- **Software developers & market analysts**: identify trends, market niches and opportunities to position existing and new apps.
- **Appstore providers**: stimulate sales by providing users with efficient browsing interfaces.

Research Question:
Can we get a better overview of the content of the app repository automatically?

Method

We constructed a topic model that automatically categorizes iPhone apps using their textual descriptions and compared its results with the original iTunes categories.

The backbone of our approach is **Latent Dirichlet Allocation (LDA)** algorithm used for topic modeling [1]. LDA assumes that documents, represented as bags of words, exhibit different topics in different proportions:

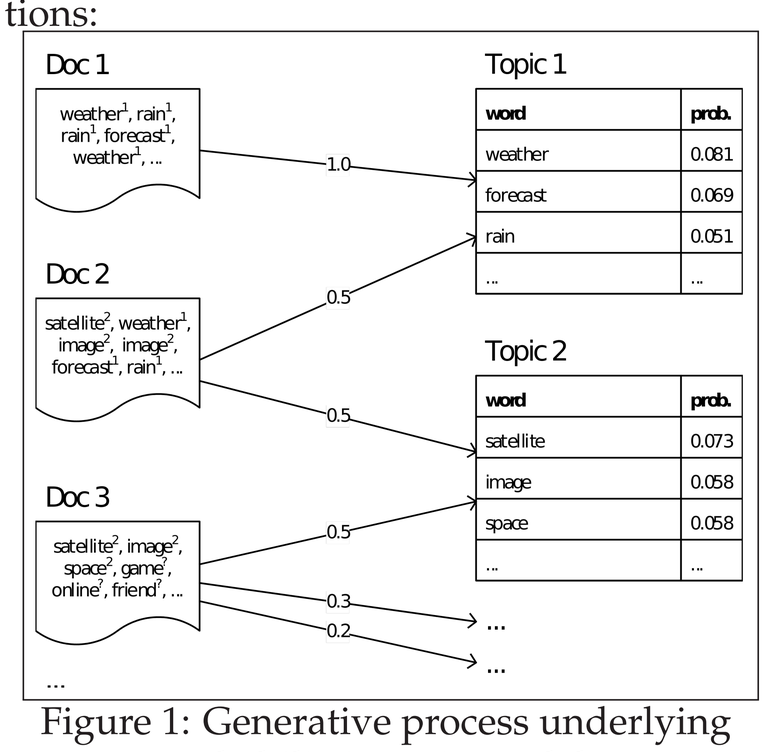

Figure 1: Generative process underlying probabilistic topic models

- **Dataset**: > 700,000 iPhone apps metadata in JSON from iTunes API (October 2013).
- **Software**: MongoDB, Google Compact Language Detector, NLTK, Gensim, Google Charts.

References

[1] Blei, D., Ng, A., and Jordan, M. 2003. "Latent dirichlet allocation," The Journal of Machine Learning Research (3), pp. 993-1022.

Results

Table 1: LDA topics and the corresponding categories

%Apps	#Apps	LDA Topic	iTunes Parent Categories	iTunes Sub-categories
9.7	83,476	0.110*game + 0.032*play + 0.024*level + 0.018*player	Games, Entertainment	Arcade, Action, etc.[5]
4.4	37,771	0.051*news + 0.037*latest + 0.023*event + 0.022*access		News
4.1	35,553	0.020*game + 0.012*enemy + 0.010*world + 0.009*battle	Games, Entertainment	Arcade, Action, Adventure, Strategy, Role Playing
4	35,015	0.019*list + 0.019*email + 0.018*note + 0.015*data	Utilities, Productivity	
3.9	33,360	0.022*business + 0.016*product + 0.013*service + 0.012*information	Business	Professional & Trade, Business & Investing
3.8	32,731	0.061*child + 0.038*kid + 0.020*story + 0.016*learn	Education, Games	Educational
3.6	31,327	0.034*like + 0.026*want + 0.022*know + 0.019*love	Entertainment	
3.6	30,935	0.024*help + 0.023*time + 0.019*make + 0.018*need		
3.5	29,697	0.122*photo + 0.042*picture + 0.031*image + 0.029*camera	Entertainment	Photo & Video
3.4	29,330	0.024*screen + 0.024*iphone + 0.020*text + 0.018*touch	Entertainment	
3.3	28,323	0.047*search + 0.037*find + 0.028*event + 0.026*view	Lifestyle	
2.6	22,126	0.071*video + 0.054*share + 0.048*friend + 0.047*facebook	Social Networking, Entertainment	
2.6	22,022	0.054*application + 0.053*user + 0.041*iphone + 0.018*ipad	Business	
2.4	20,946	0.054*time + 0.039*button + 0.016*screen + 0.015*timer	Utilities	
1.9	16,499	0.022*learn + 0.016*video + 0.015*student + 0.013*learning	Education	

[5] 14 out of 18 subcategories of *Games* (excluding *Educational*, *Music*, *Role Playing* and *Sports*).

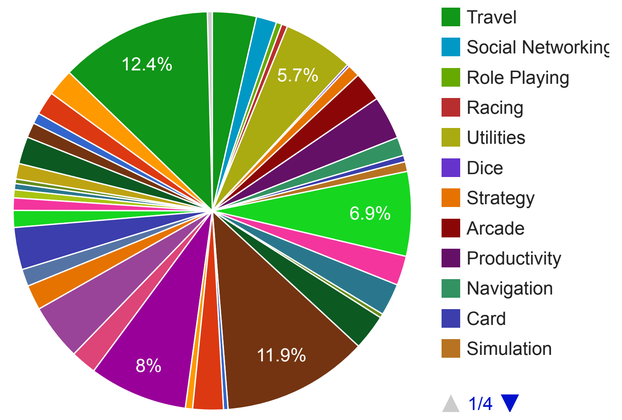

Figure 1: iTunes categories

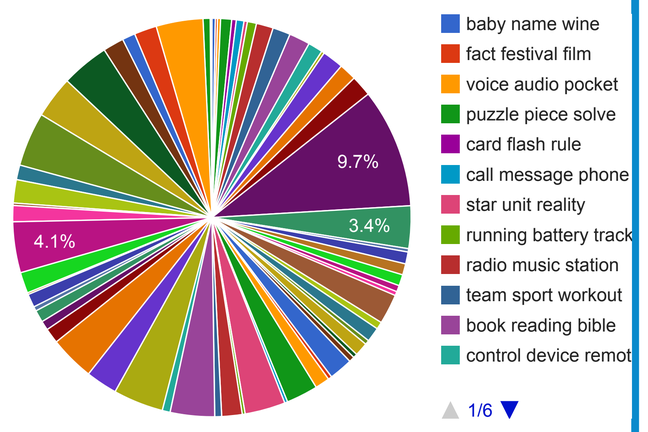

Figure 2: LDA topics

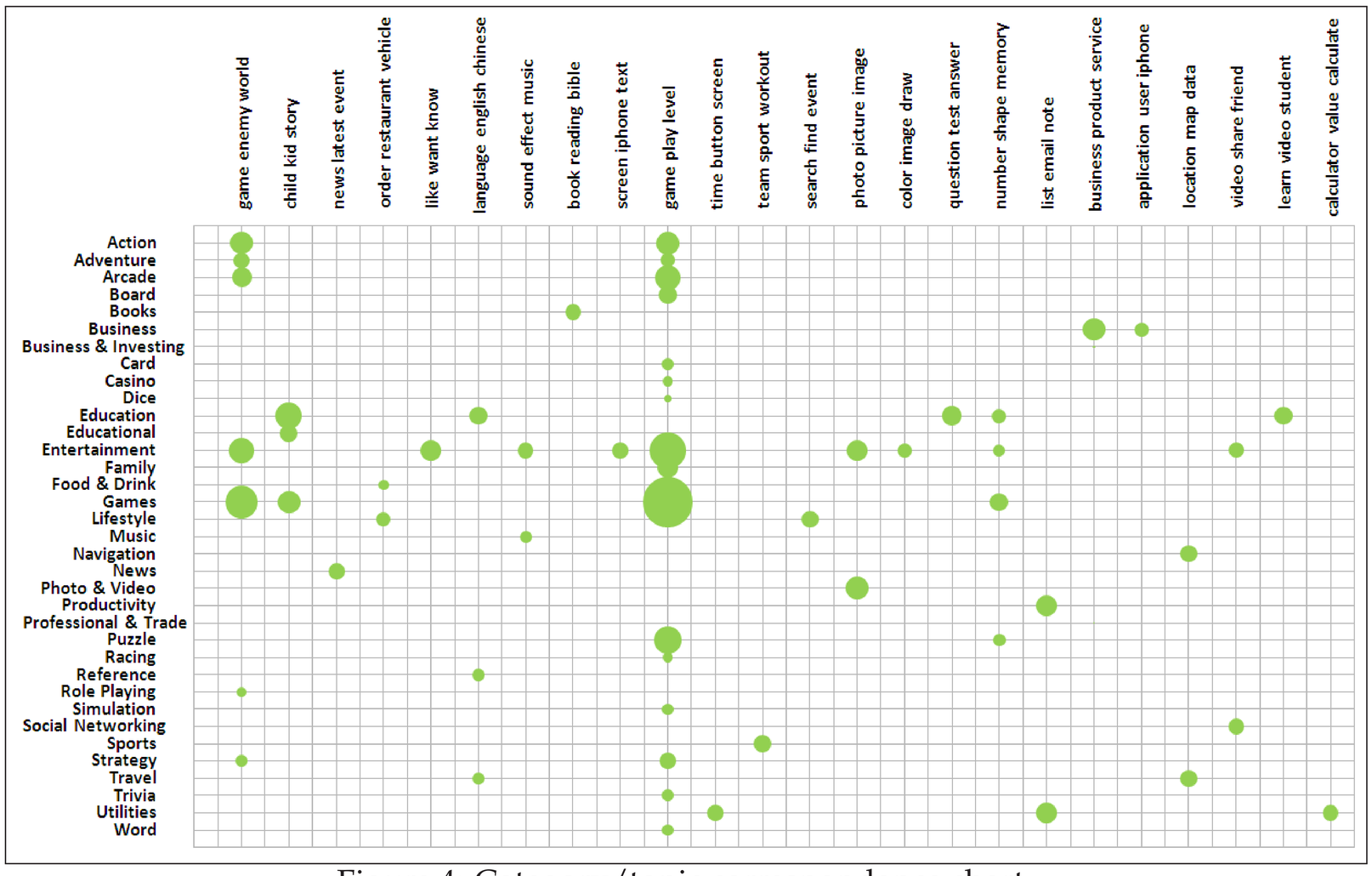

Figure 4: Category/topic correspondance chart

Future Directions

- Improve our approach towards building a **hierarchy of topics** (e.g. Hierarchical LDA).
- **Evaluate** quality of classification schemes.
- Conduct **correlation analysis** between the various categories and other metadata in order to spot interesting patterns and dependencies.
- **Detect trends** describing evolution of the market place over time.
- Extend analysis to other repositories (e.g. **Google Play Store**) to compare the offer across the major app stores.

Summary

- LDA produces more balanced classes.
- Topics identify potential subcategories.
- High loaded terms reveal semantics of the category and frequent patterns.
- Word probabilities are very low.

Feedback

The interactive visualizations and other supplementary materials are available at `https://github.com/vendi12/whatsIn`

Contact: **svitlana.vakulenko@modul.ac.at**