MINING ANALOGICAL LIBRARIES IN Q&A DISCUSSIONS

— INCORPORATING RELATIONAL & CATEGORICAL KNOWLEDGE INTO WORD EMBEDDING

CHUNYANG CHEN, SA GAO, ZHENCHANG XING

School of Computer, Nanyang Technological University, Singapore

Analogical Technique Questions

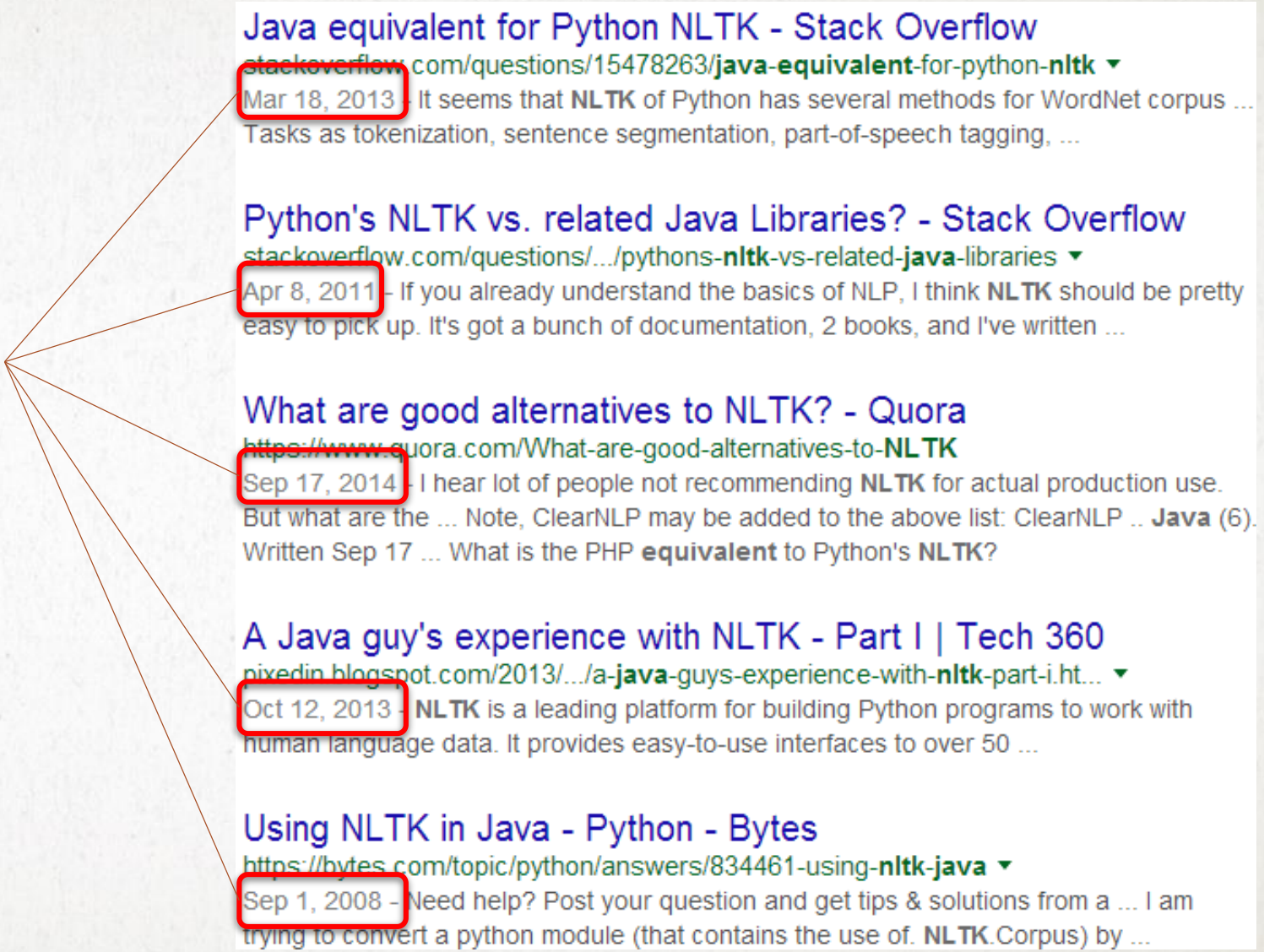
- ? is to A as b is to B?
  - e.g., What are Java's equivalent libraries to Python's NLTK ?

- When do developers ask it?
  - They are not satisfied with current libraries;
  - They migrate from one language to another.

Analogical Technique Questions

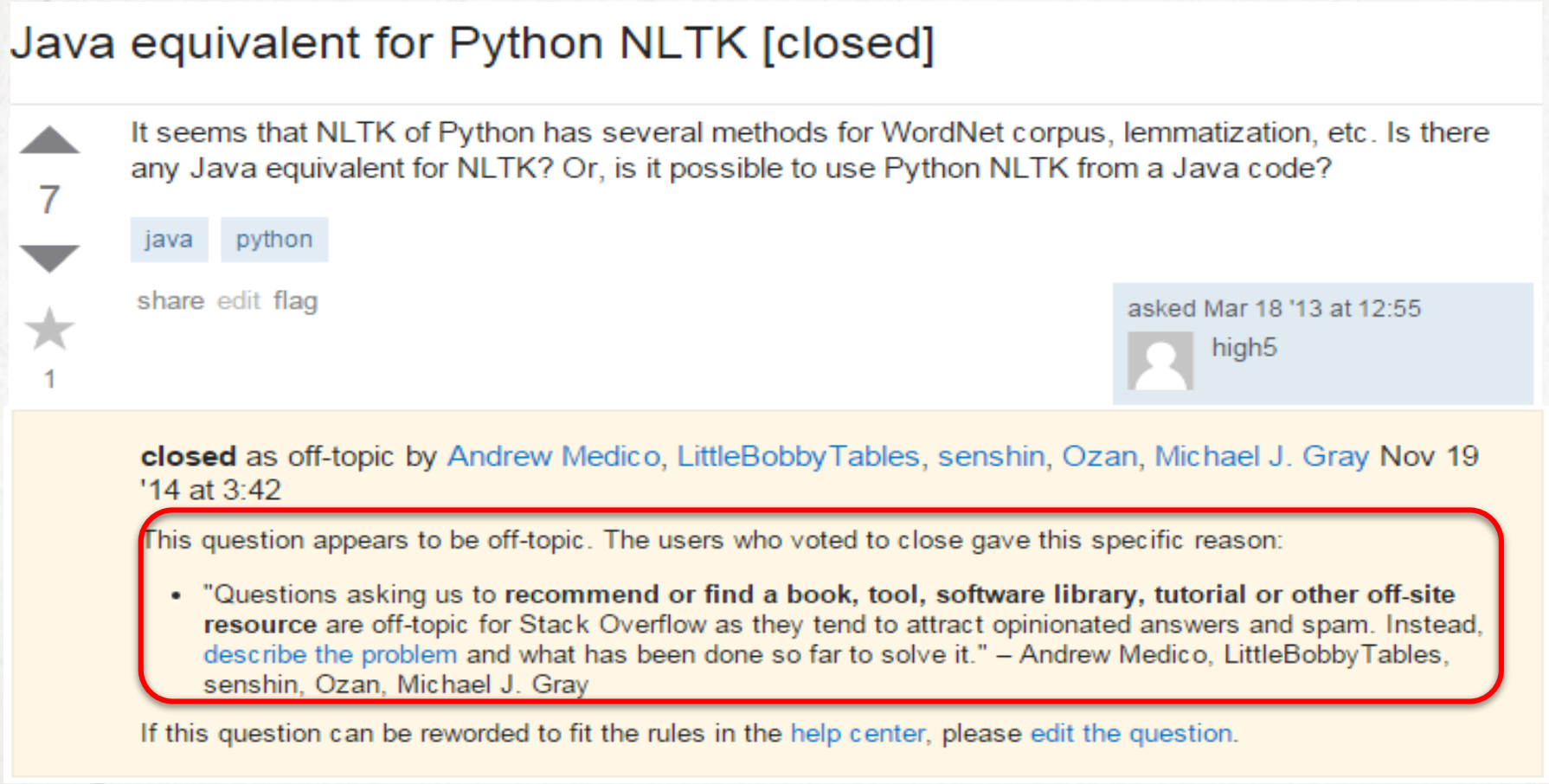
- When we ask in Google:

Out of date & subjective



Analogical Technique Questions

- When we ask it in Q&A site in Stack Overflow
  - Not allowed in Stack Overflow as it is too **subjective**.



Java equivalent for Python NLTK [closed]

7

It seems that NLTK of Python has several methods for WordNet corpus, lemmatization, etc. Is there any Java equivalent for NLTK? Or, is it possible to use Python NLTK from a Java code?

java python

share edit flag

asked Mar 18 '13 at 12:55

high5

1

closed as off-topic by Andrew Medico, LittleBobbyTables, senshin, Ozan, Michael J. Gray Nov 19 '14 at 3:42

This question appears to be off-topic. The users who voted to close gave this specific reason:

- "Questions asking us to **recommend or find a book, tool, software library, tutorial or other off-site resource** are off-topic for Stack Overflow as they tend to attract opinionated answers and spam. Instead, describe the problem and what has been done so far to solve it." – Andrew Medico, LittleBobbyTables, senshin, Ozan, Michael J. Gray

If this question can be reworded to fit the rules in the help center, please edit the question.

http://stackoverflow.com/questions/15478263/java-equivalent-for-python-nltk

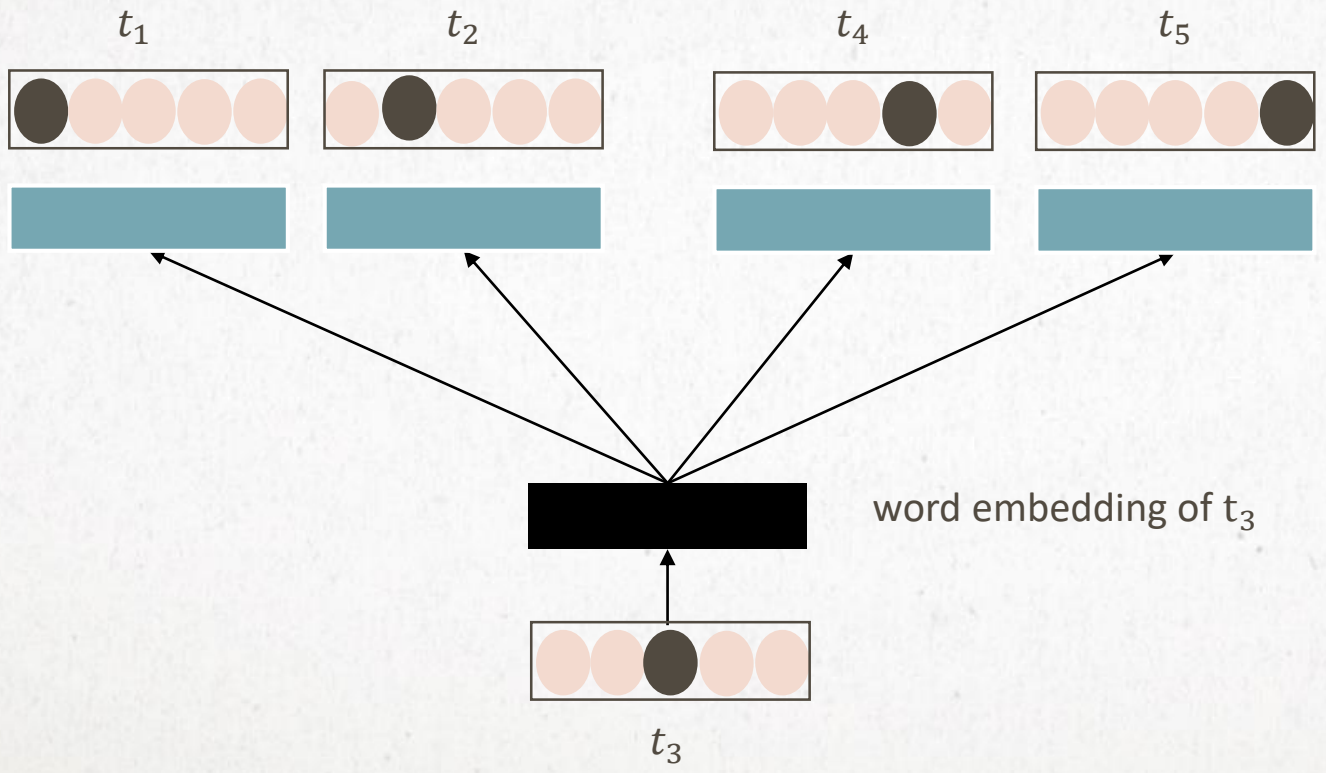
Can we answer such questions automatically and objectively?

A Key Observation

- Analogical libraries often appear in **similar context**.

Continuous Skip-Gram Model

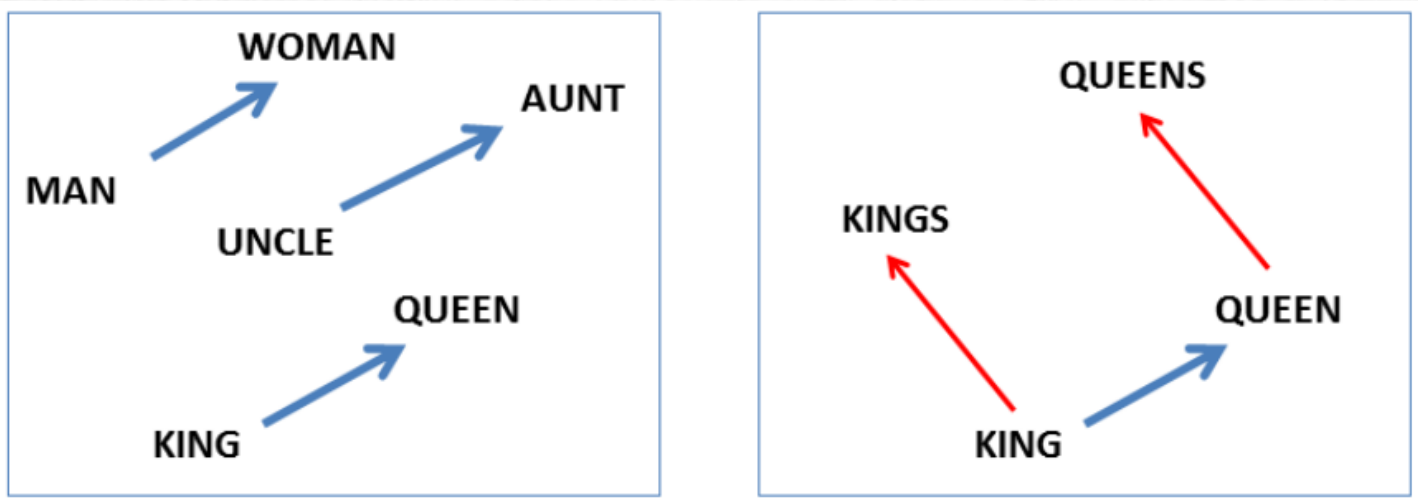
- Learn word vector representations that are good at predicting the nearby words i.e., word to vector (w2v).



Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality." *Advances in neural information processing systems*. 2013.

Relational Similarity

- Word embedding encodes relation between pairs of words:
  - E.g., Queen – King ~ Woman – Man (analogy reasoning)



- Solve analogy question "? is to A as b is to B" by:

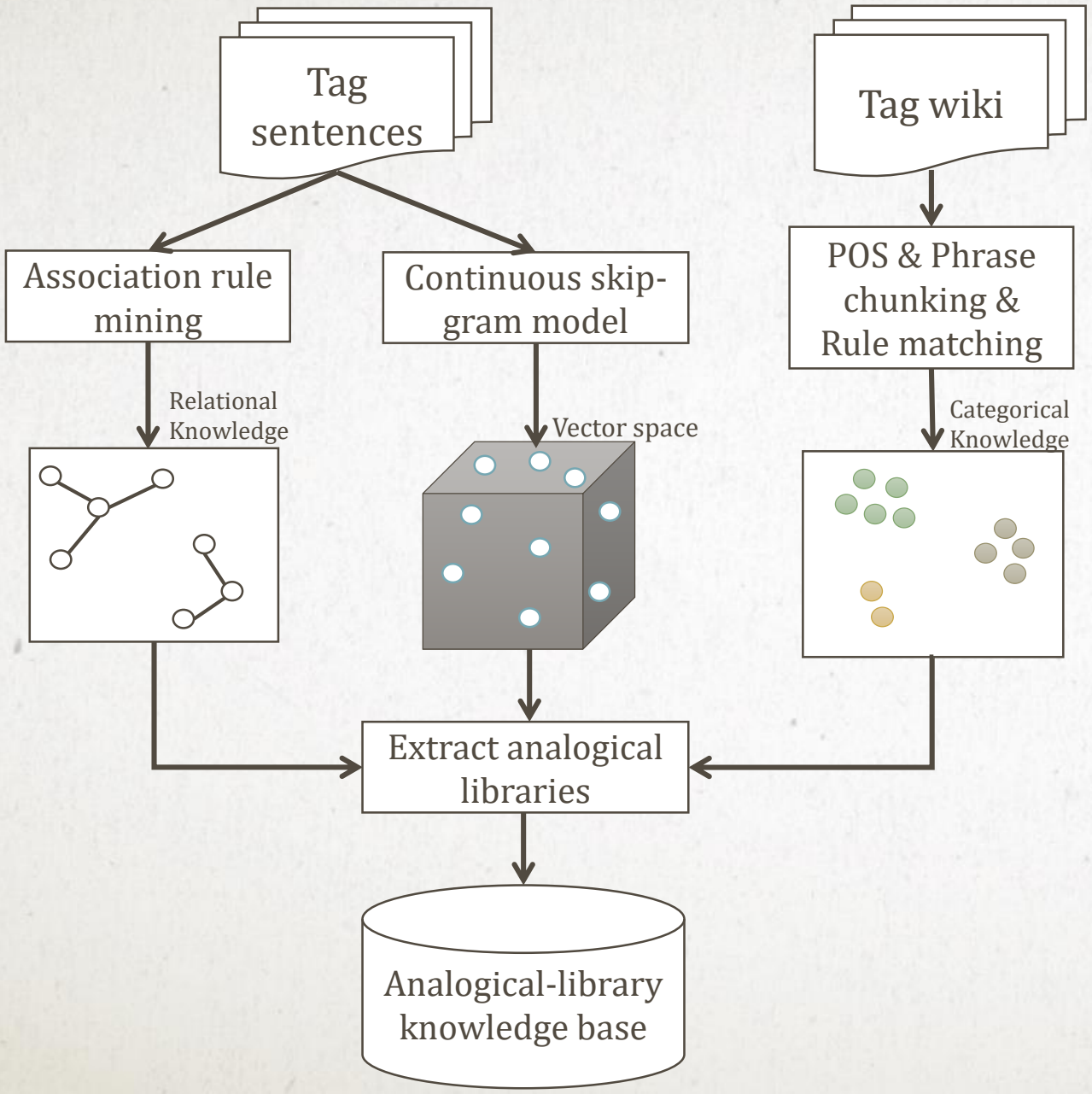
$$\arg\max_{a \in V} \cos(a, A - B + b)$$

Apply the model to our task?

- Original word embedding is designed for general text;
  - Tag sentences are too short ! (<= 5);
  - No linguistics in tag sentence !

- With solely original word embedding:
  - Too many **false positives,** so we need to **customize** it !

Overall Method



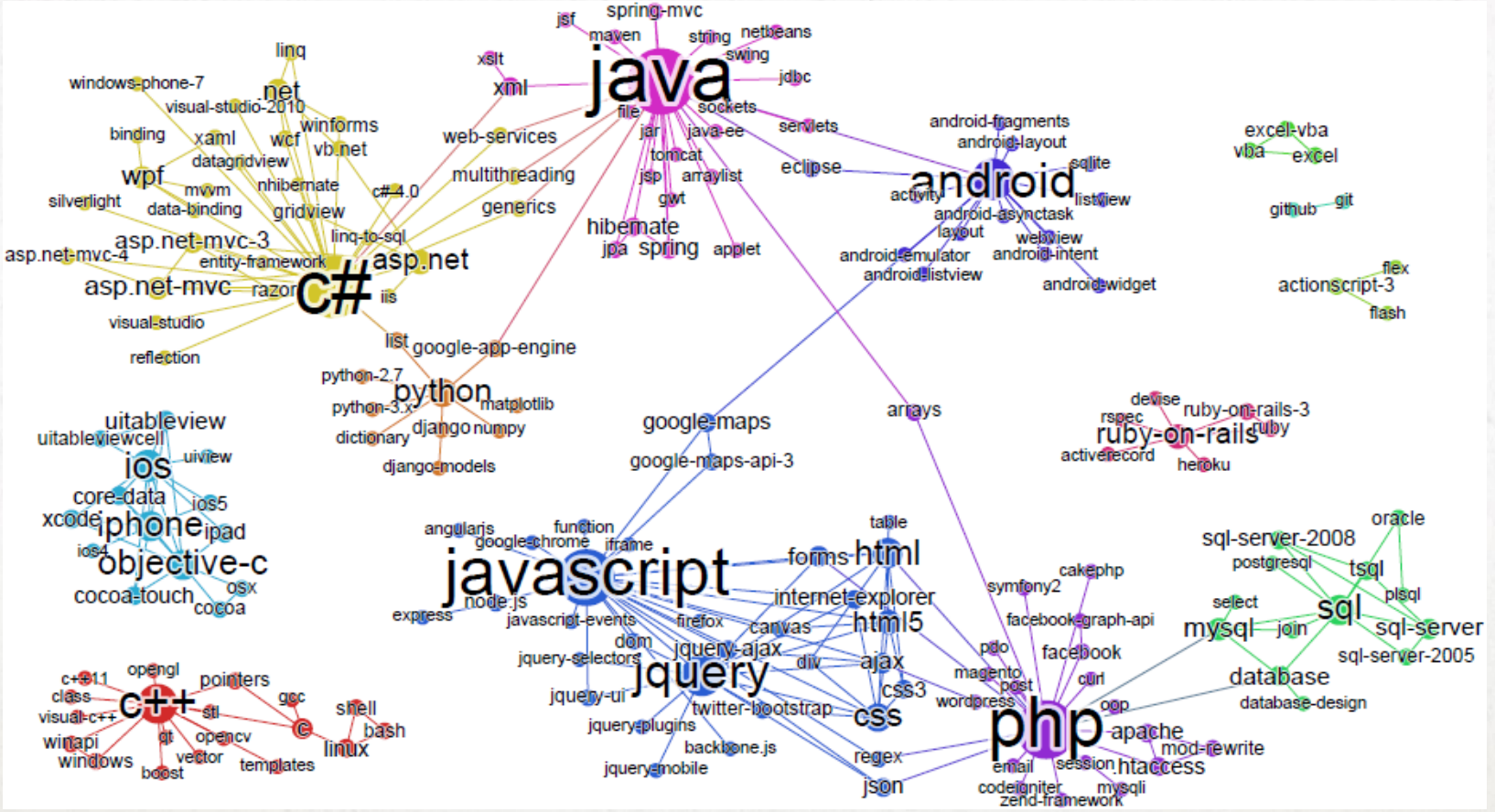
Not only semantic knowledge learned by word embedding, but also incorporate:

- Relational Knowledge;
- Categorical Knowledge.

Mining Relational Knowledge

- Get tag relations by association rule mining:
  - E.g., java → opennlp, python → nltk
- Link rules → a technology associative network.

Part of the graph



Chen, Chunyang, and Zhenchang Xing. "Mining technology landscape from stack overflow." In *Proceedings of the 10th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement*, p. 14. ACM, 2016.

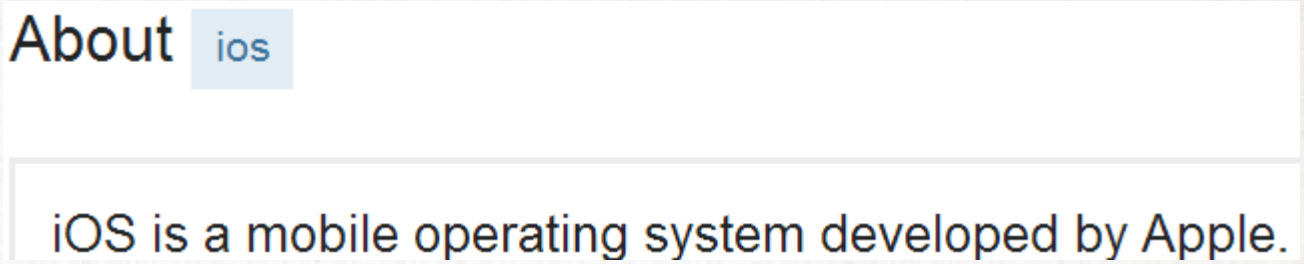
Categorical Knowledge

- Tags in Stack Overflow belong to different categories:
  - Java → programming language;
  - Eclipse → IDE;
  - Binary-search → algorithm;
  - Caching → mechanism;
  - D3.js → library.

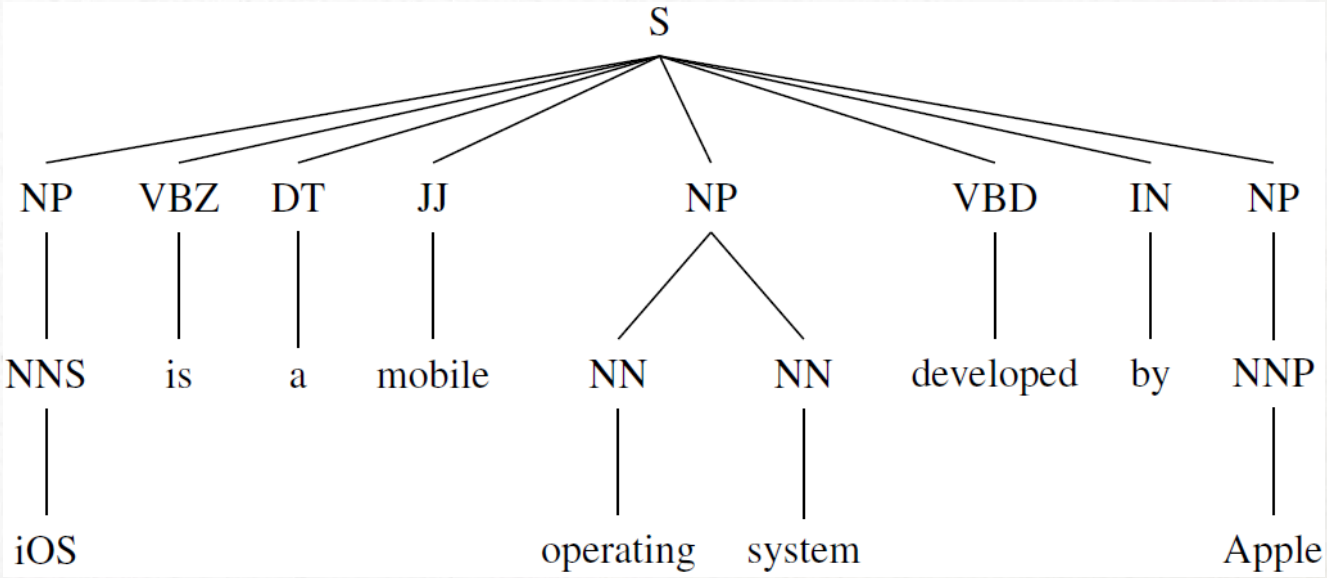
- Original word embedding doesn't know that we need library tags in this work !

Mining Categorical Knowledge

- Community TagWiki info
  - First sentence



- Get the category of the tag
  - Part-of-Speech tagging and phrase chunking
  - First noun phrase after "be" is category label

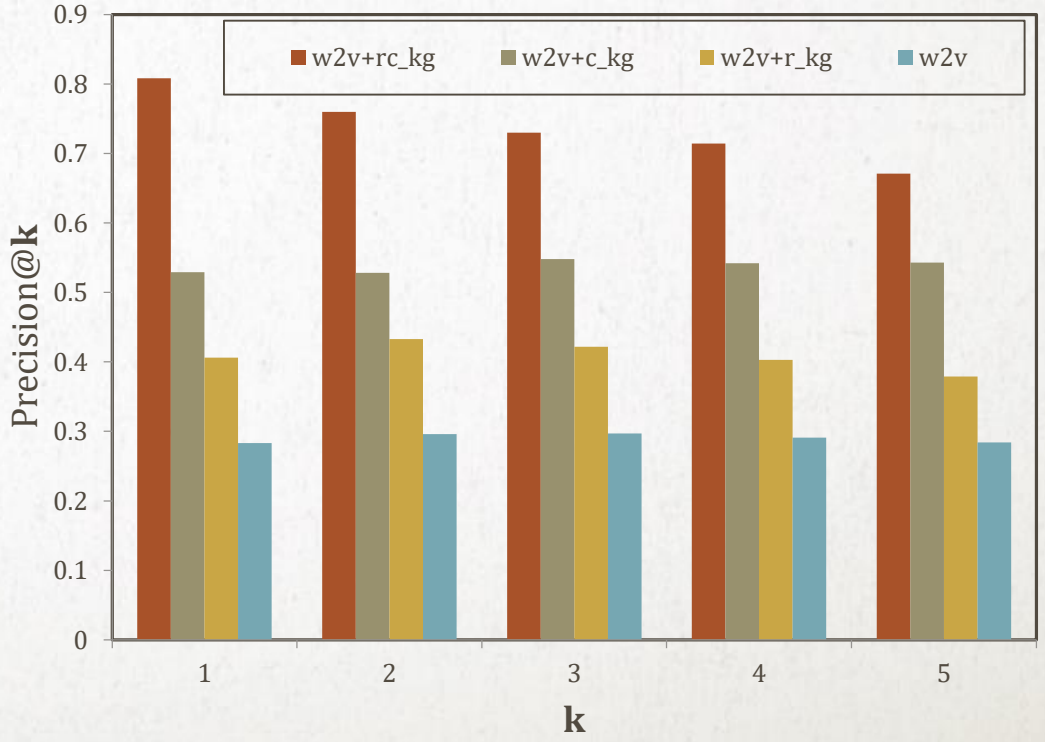


http://stackoverflow.com/tags/ios/info

Evaluation

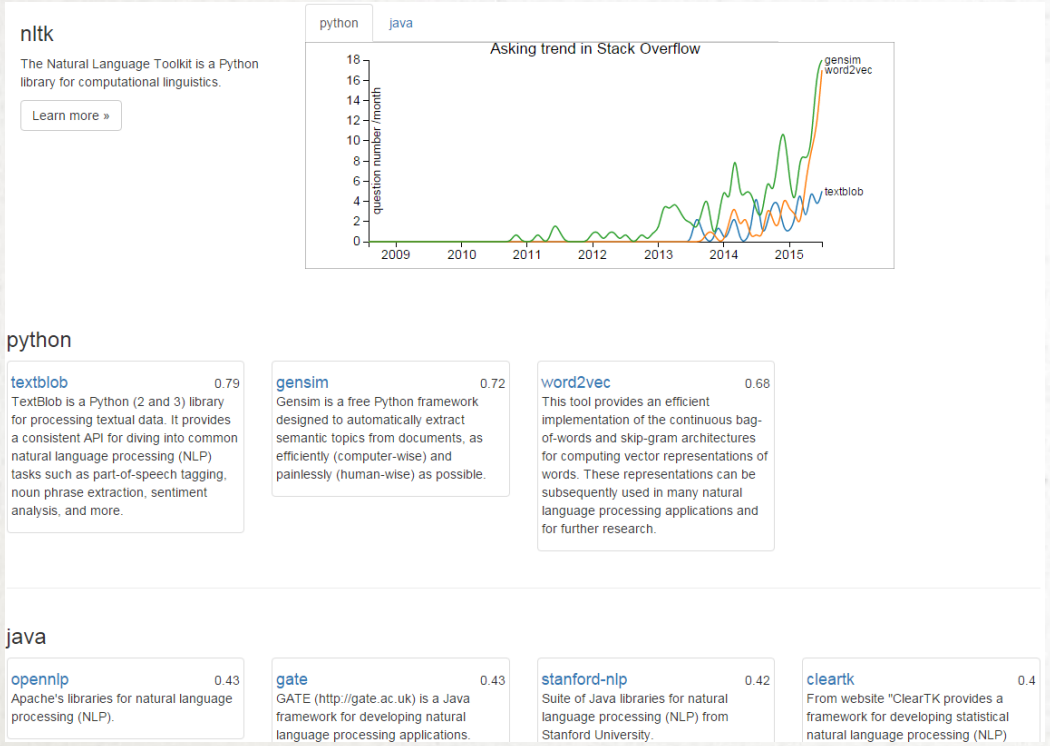
- Dataset
  - 7 years Stack Overflow data (07/2008 - 08/2015)
  - 9.97 million questions, 36,197 tags with TagWiki and 7,783 library tags

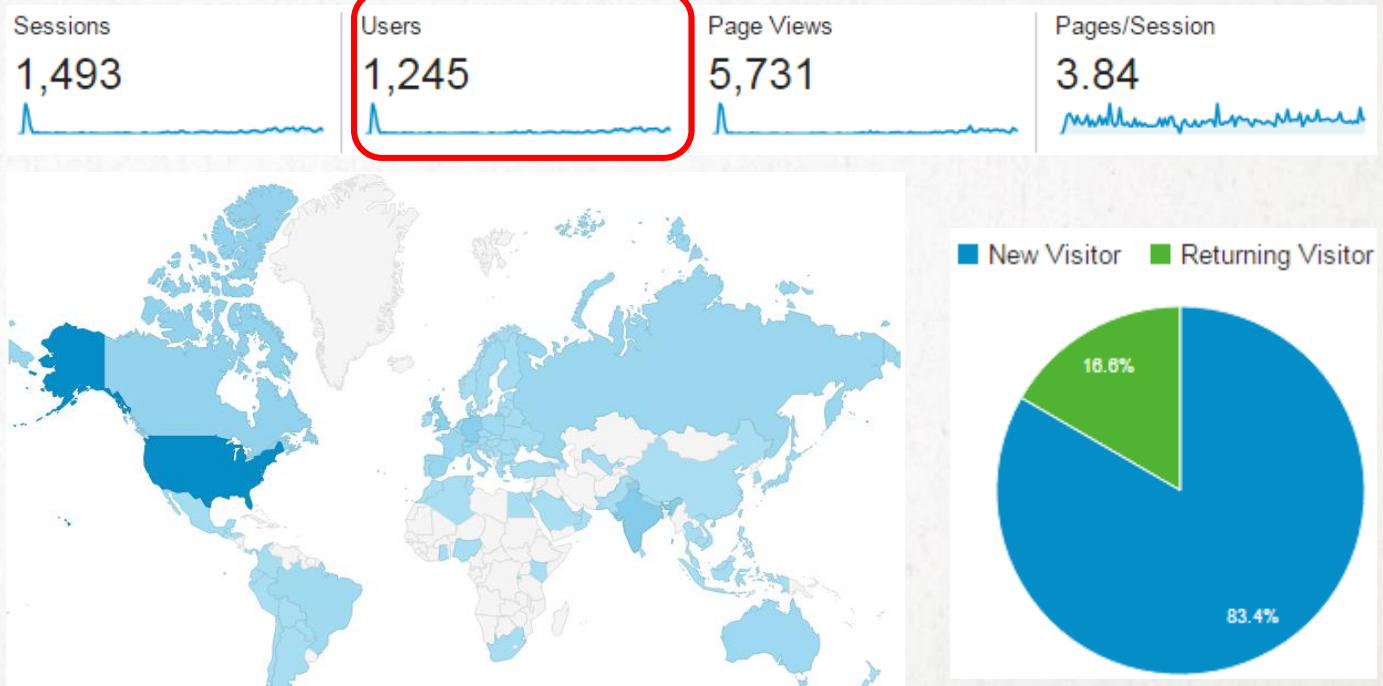
- Evaluate three components in the approach:
  - The accuracy of tag categorization
    - 84% are correctly categorized
  - The semantic distance of tag correlations
    - 88% appear in Google Trends
  - The recommendation of analogical libraries
    - Precision@1 = 0.81, Precision@5 = 0.67



Our Website & Visiting Statistic

https://graphofknowledge.appspot.com/similarTech





THANKS A LOT FOR THE LISTENING

- Chen, Chunyang, Sa Gao, and Zhenchang Xing. "Mining analogical libraries in q&a discussions--incorporating relational and categorical knowledge into word embedding." In *2016 IEEE 23rd international conference on software analysis, evolution, and reengineering (SANER)*, vol. 1, pp. 338-348. IEEE, 2016.
- Chen, Chunyang, and Zhenchang Xing. "Similartech: automatically recommend analogical libraries across different programming languages." In *2016 31st IEEE/ACM International Conference on Automated Software Engineering (ASE)*, pp. 834-839. IEEE, 2016.
- Chen, Chunyang, Zhenchang Xing, and Yang Liu. "What's spain's paris? mining analogical libraries from q&a discussions." *Empirical Software Engineering* 24, no. 3 (2019): 1155-1194.