# Bank Q&A System

Group 42
Yuewei Na, Naiqing Song, Xiang Zhang, Shurui Liu

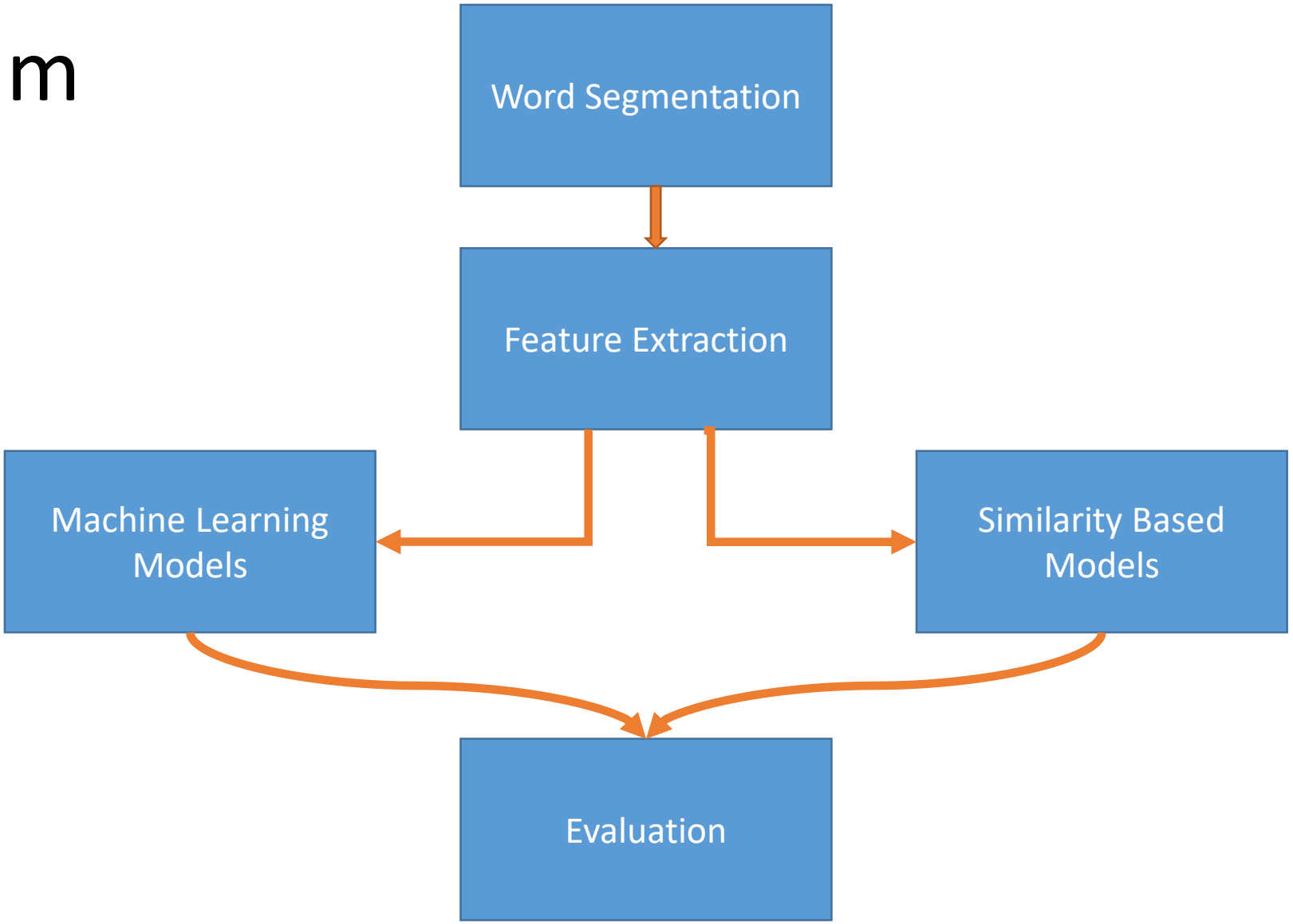
# Background

- A Chinese bank is looking for an online Q&A system to help customer service department.
- It used to be like:
  - Customers calling in, other customers have to wait on the line for a long time.
- What we want:
  - Instead of calling in, customers can go to website and type in the question to get answers quickly.

# What we have

- A corpus that contains all the possible questions and their extensions.
- For example,
  - Original question (translation): What type of ID do I need to have when I open up an account?
  - Extended questions could be
    - Can I use my passport as a valid ID when I open up an account?
    - Can I use my student ID/employee ID when I open up an account?
    - What can be considered as valid ID for foreign people when open up an account?
    - What ID can be considered as valid ID when I open up an account?

# Diagram


Word Segmentation
Feature Extraction
Machine Learning Models
Similarity Based Models
Evaluation

# Segmentation Approaches

The group found four different tools to help with segmentation on Chinese.

- Stanford
- THULAC
- ICTCLAS
- Jieba

Use information extraction related techniques to retrieve keywords and find related questions from the corpus.

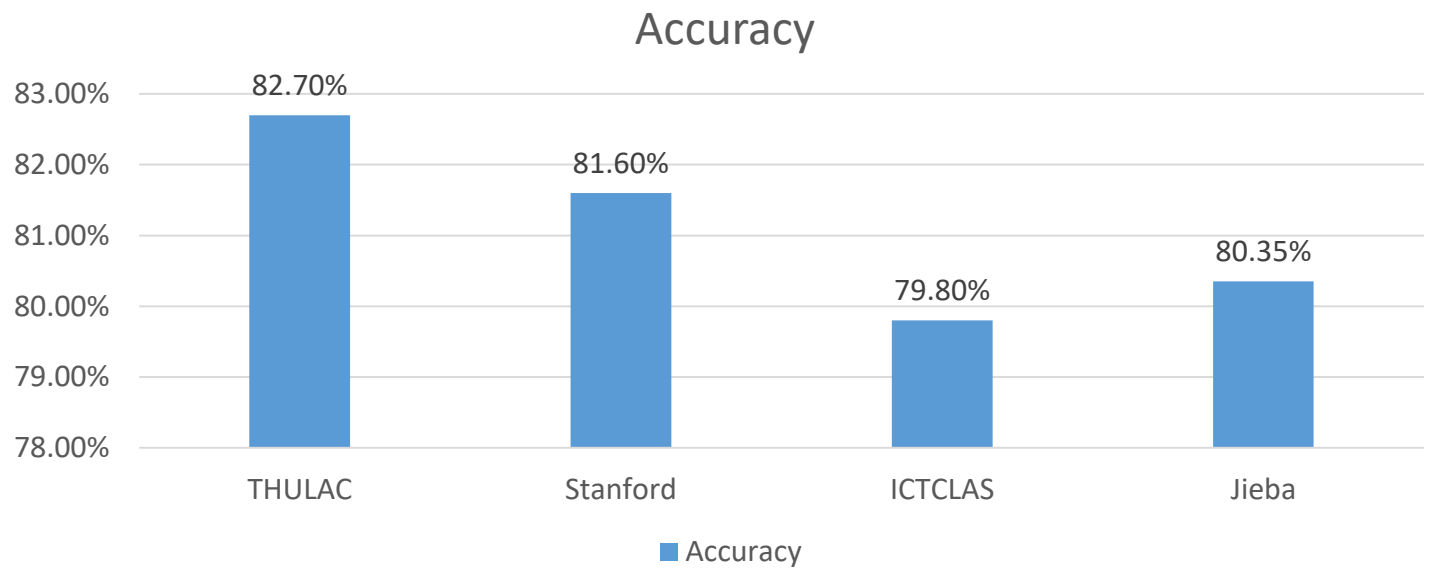
# Segment Tool Comparison cont.

Eg. 零存整取不足半年计息

- Stanford: 零存/ 整取/ 不足/ 半/ 年/ 计息/
- THULAC: 零存/ 整取/ 不足/ 半/ 年/ 计/ 息
- ICTCLAS: 零/ 存/ 整取/ 不足/ 半年/ 计息
- Jieba: 零存整取/不足/半年/计息

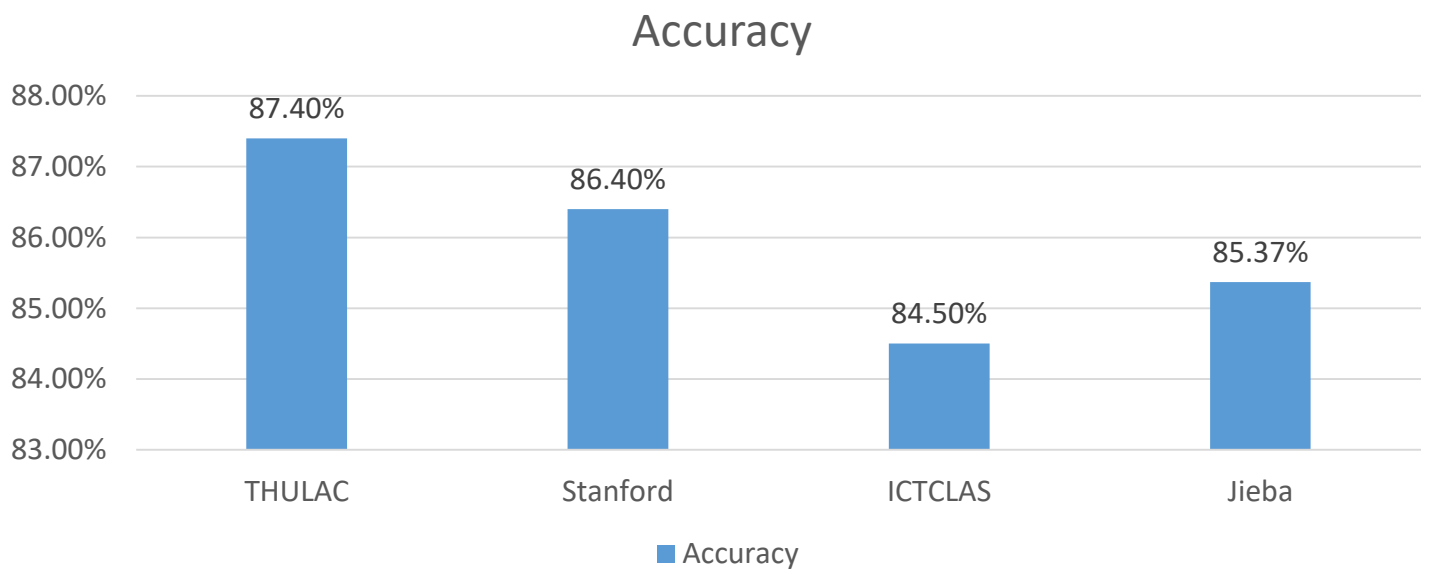
# Classification--Nearest Neighbors

- Feature: bag of word
- Parameter: 5-fold cross validation
- Toolkit: sklearn
- Accuracy



# Classification--Random Forest

- Feature: bag of word
- Parameter: 0.816
- Toolkit: sklearn
- Accuracy



# Similarity Based Method

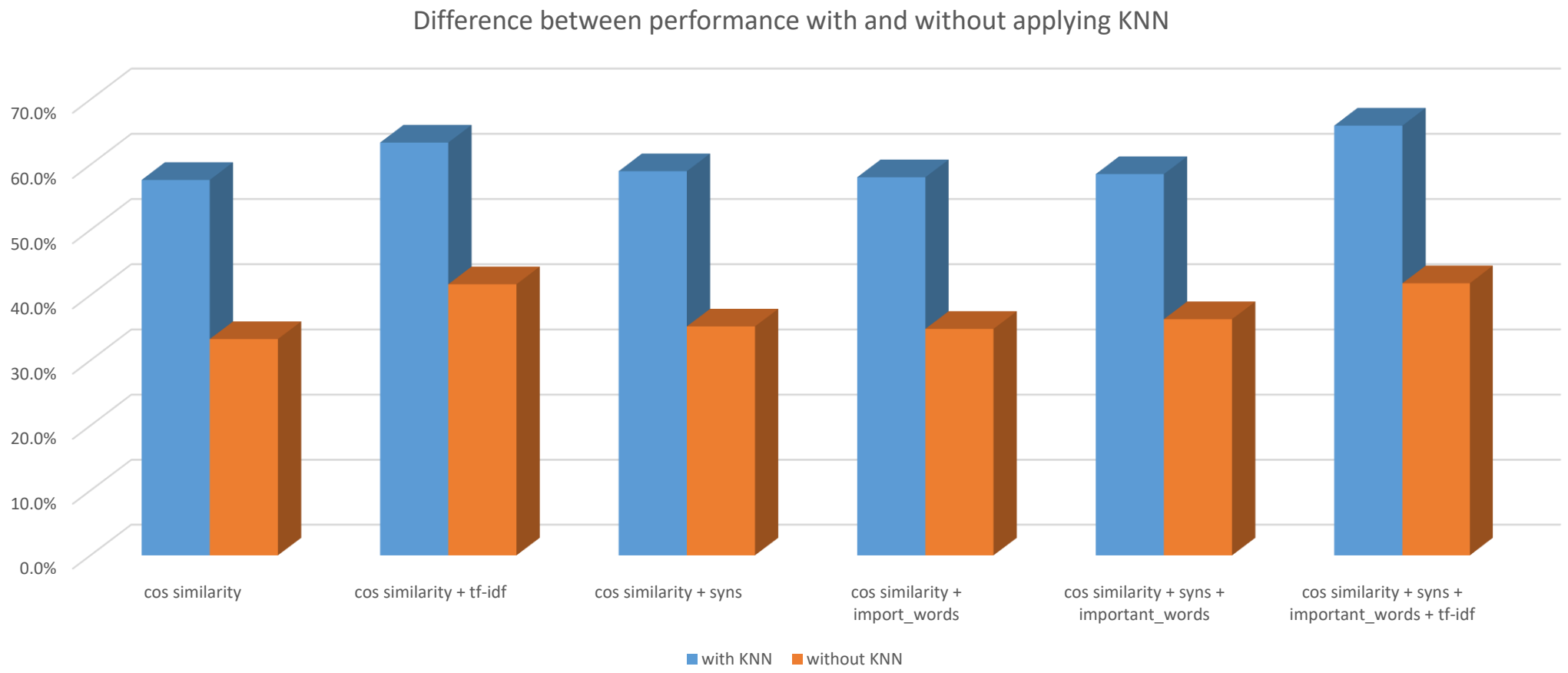
1. Cosine similarity: computing the similarity of query and answer based on words co-concurrence

2. Computing the similarity between not only query and answer, but query and queries labelled by the answer.

3. Considering word similarity rather than only word co-concurrence

4. Considering weights of word by tf-idf and way of human labelling

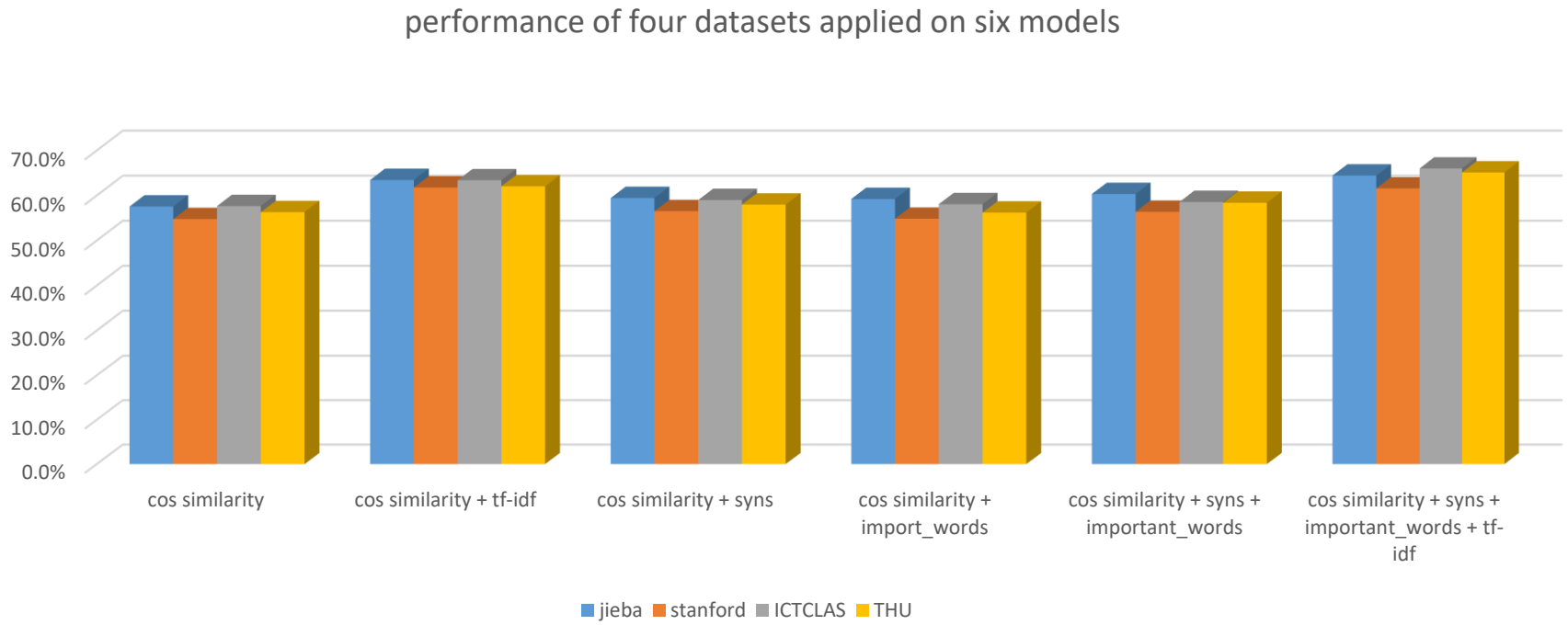
# Difference between performances with and without considering similar queries as mentioned



# Compare the performances of different methods as mentioned


Conclusion: The models combined with tf-idf and combing all are better than the others

# We used four tokenizing tools and compared their performances




Conclusion: The performance of dataset tokenized by ICTCLAS and jieba are better than the other twos

# Questions?