StackOverflowTagger

By Dmitrii Stoianov,
Elena Cheprasova,
Ilya Gulyaev,
Yaroslav Rogov

Problem statement

80% of data generated each day is unstructured

extremely difficult to analyze and process

automated keyword extraction:

- summarize a text
- index data

Problem statement - Stack Overflow

Tagging helps in finding users, that can answer a question

Tag may filled manually by user, but it may be missed or used incorrect - less probability to have a correct answer

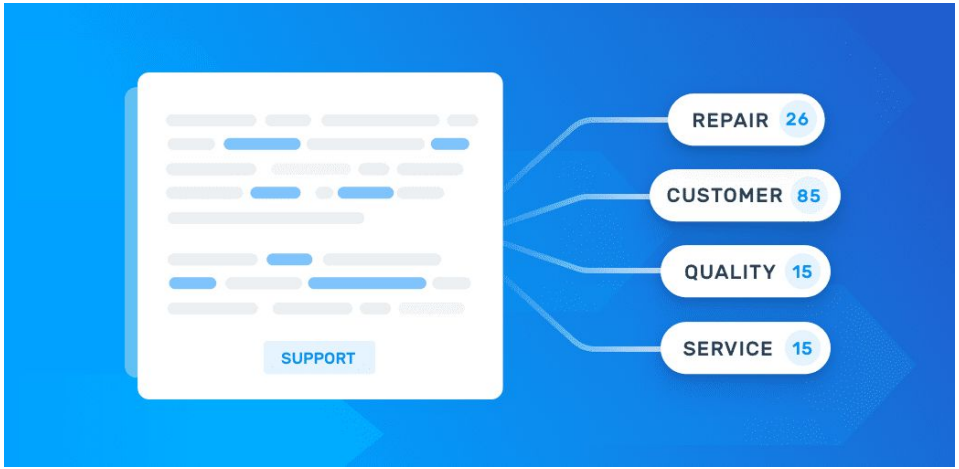
Accurate automatic tagging by title and body of question can solve the problem

Better ecosystem - direct business impact

Tagging: Existing Solutions

Statistical methods:

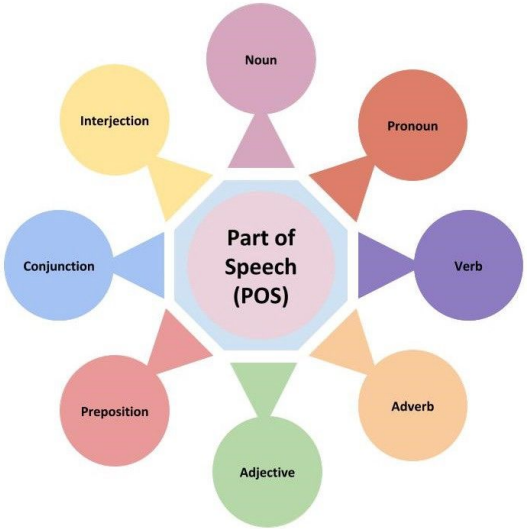
- Word Frequency
- Word Collocations
- TD-IDF
- RATE



Tagging: Existing Solutions

Linguistic methods:

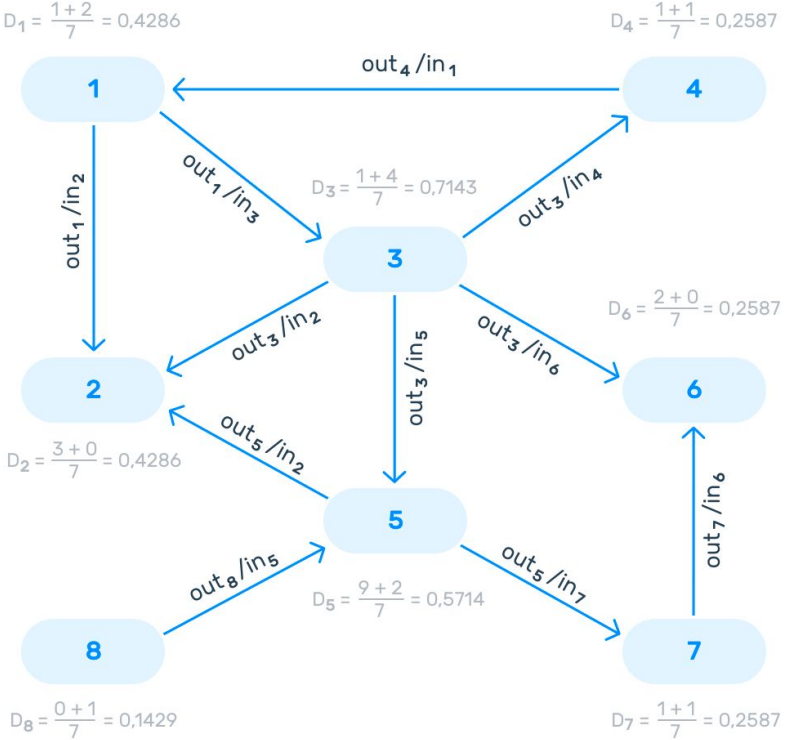
- Part Of Speech tagging
- Grammar dependency
- Selection on Informative features (bold, italic, font size)



Tagging: Existing Solutions

Graph-based methods:

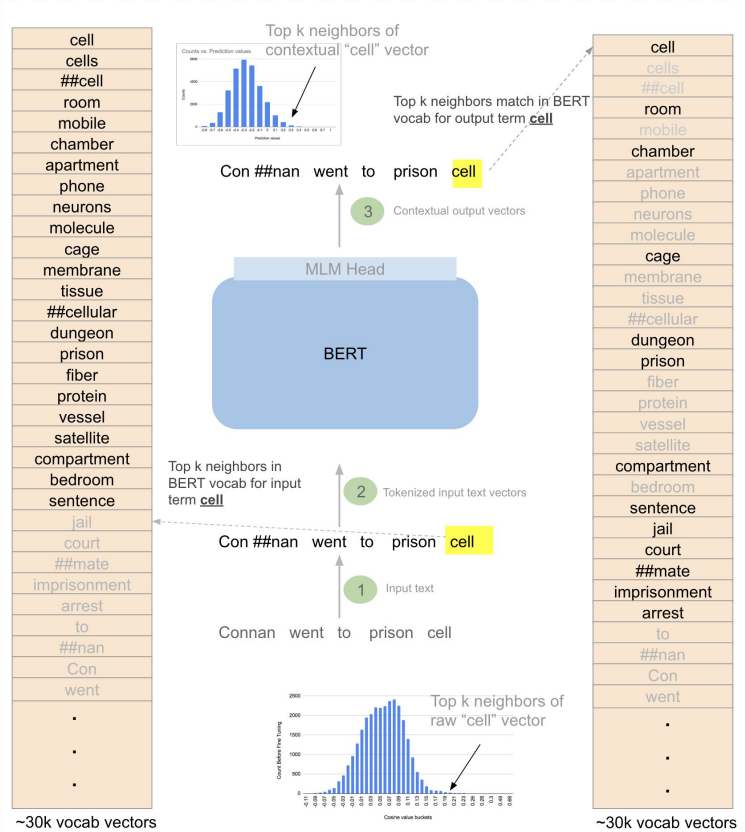
- TextRank Model



Tagging: Existing Solutions

ML methods to extract keywords:

- SVM
- Deep Learning
- BERT (transformers)



Tagging: Existing Solutions

ML methods for multi-label classification:

- Logistic Regression
- SVM
- Random Forest

Implementation: Assumption #1 — Results

———

Assumption: Tags are similar to keywords

Sentence embedding with BERT model (SentenceTransformer)

Results:

>> Expected keywords: ['wordpress', 'r', 'blogs']

>> Predicted keywords: ['wordpress', 'mediawiki', 'wiki', 'blog', 'blogging']

precision – 0.14, recall – 0.23, f1 – 0.17

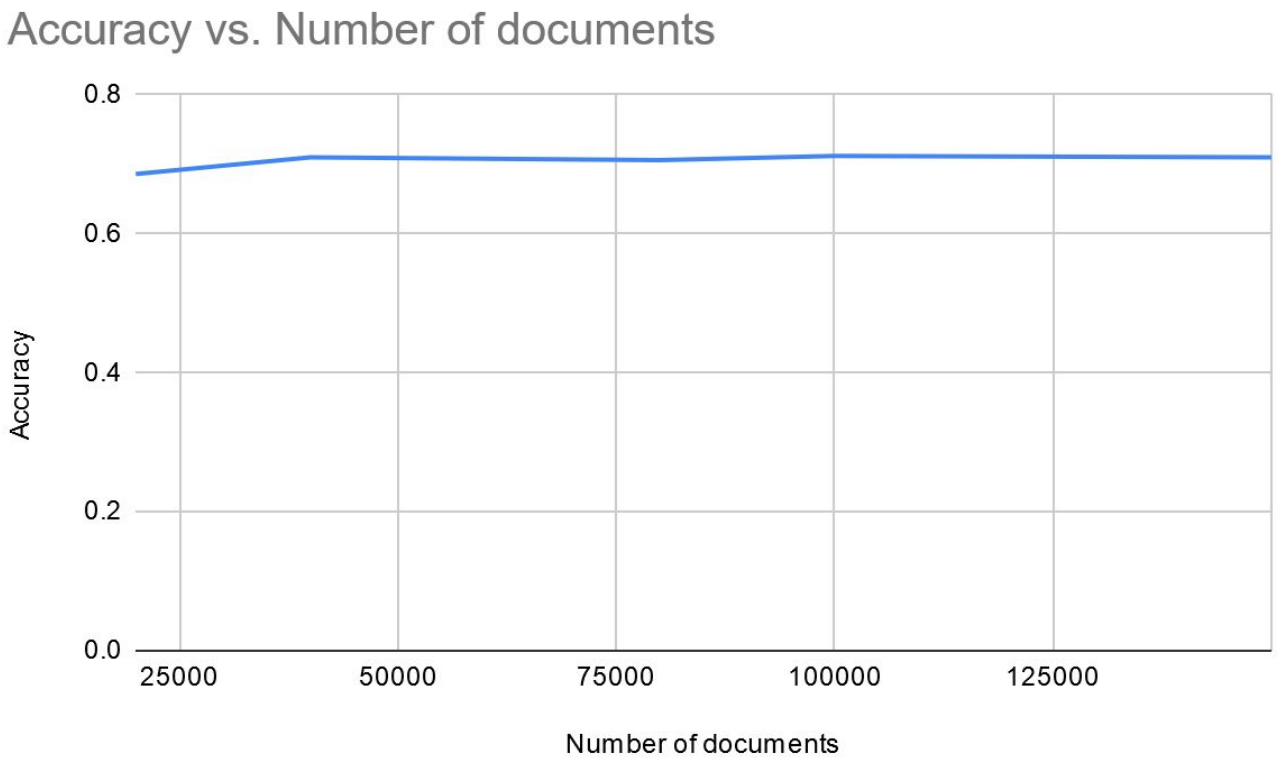
Implementation: Assumption #2

Tags can be treated like features of multiclass model

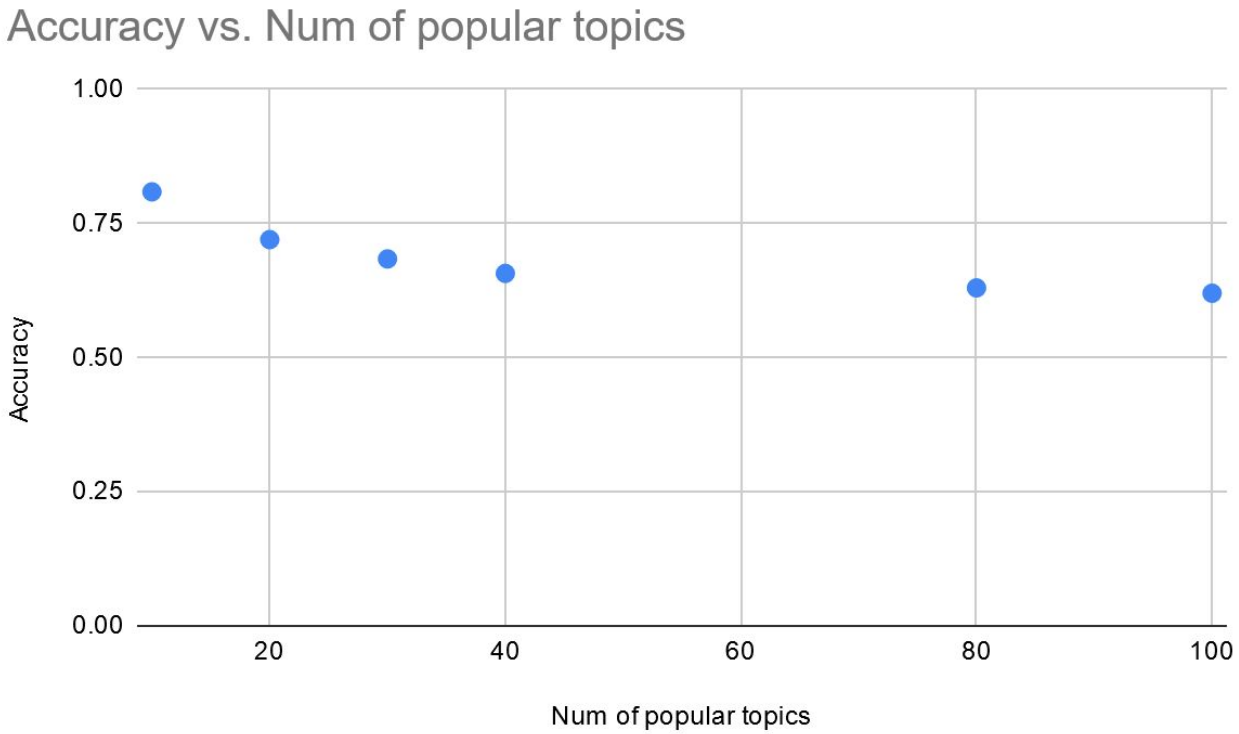
Problem:

- Big dataset is slow to process
- Number of tags has impact on accuracy

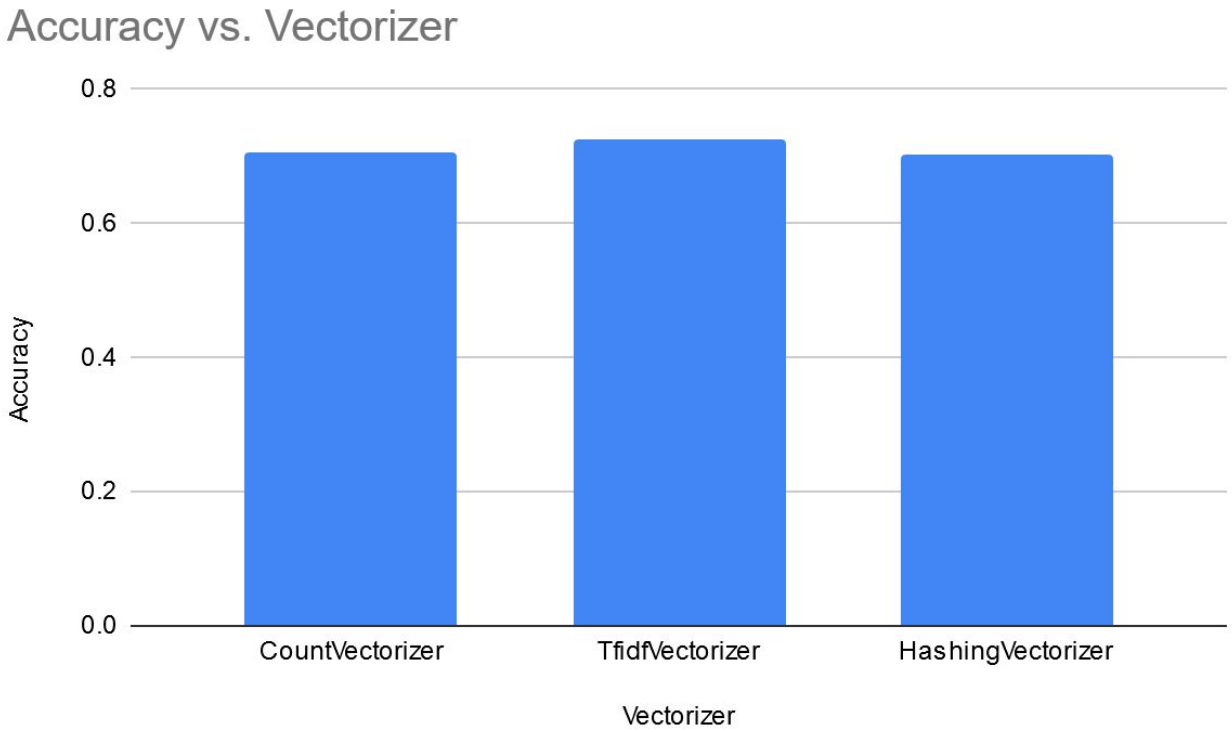
Implementation: Assumption #2 — Documents



Implementation: Assumption #2 — Features



Implementation: Assumption #2 — Results



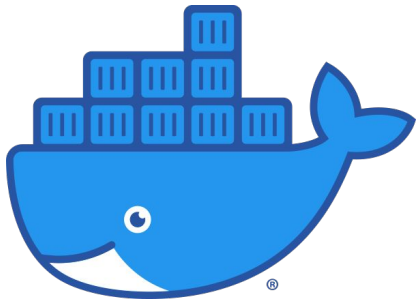
Implementation: API + Docker

API:

- Trained model wrapped by Flask based REST API
- Data preparation is the same as used in training

Docker:

- Trained model with REST API and required libraries packed to docker image
- Ready to run in any environment with docker support



Thank you for your attention!