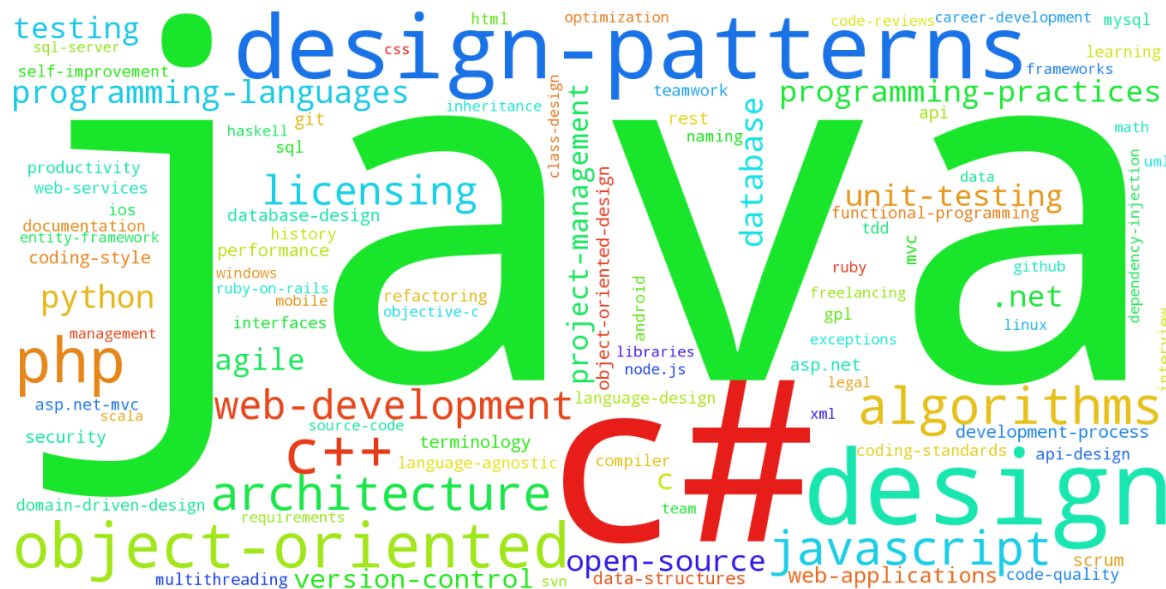


Project 4: Machine Learning

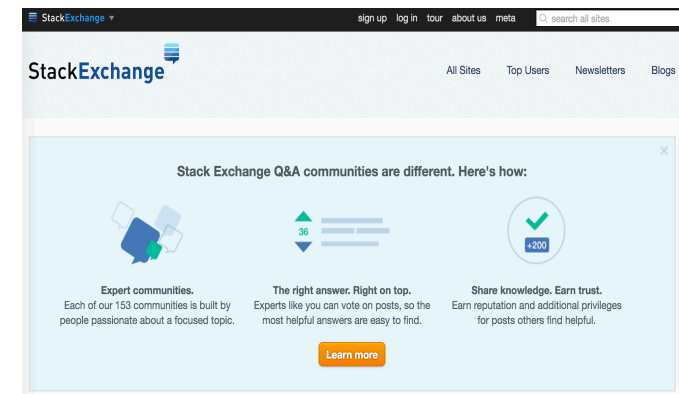
Team #7: Michael Herold, Ralph Samer



Motivation

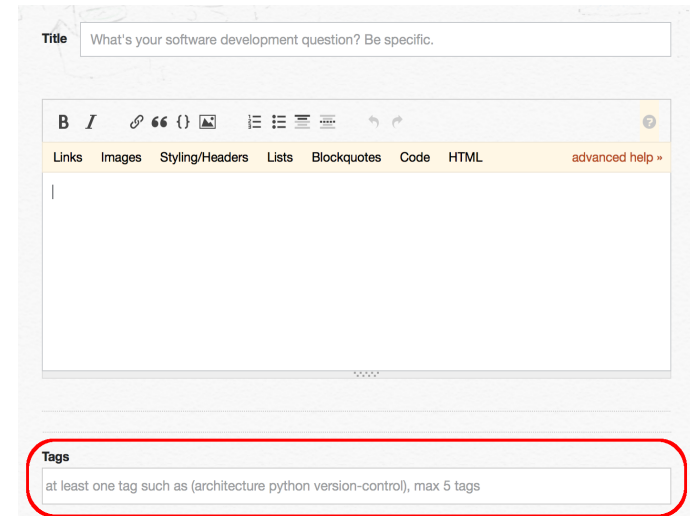
Stack Exchange:

- Experts can subscribe to tags
- Mandatory to assign proper tags to questions
- Users can create new or use existing tags
- Manual tagging may be challenging (e.g. “.Net” and “DotNet”)
- Suggesting tags for new questions is crucial



Task

- Given: new question containing a title and a body
- Goal: suggest/assign up to 5 tags from a limited list of tags
- Analyse given data set
- Supervised:
Train classifier on data set
- Unsupervised:
Find patterns in data (e.g. groups)
- Evaluate results and compare both approaches



The screenshot shows a web form for asking a question. At the top, there is a 'Title' field with the placeholder text 'What's your software development question? Be specific.' Below the title field is a rich text editor with a toolbar containing icons for bold, italic, link, unlink, quote, code, list, and other formatting options. Below the editor is a 'Tags' section, which is highlighted with a red rounded rectangle. The 'Tags' section has a label 'Tags' and a text input field with the placeholder text 'at least one tag such as (architecture python version-control), max 5 tags'.

Data Set

- Stack Exchange Data Dump from March 1st, 2016
- Subgroup: programmers.stackexchange.com
- Packed size: 179.5 MB
- Post, tags and users are stored in separate XML files
- Tags consist of 1-5 words separated by "." and "-"

```
<?xml version="1.0" encoding="utf-8"?>
<posts>
  <row Id="1" PostTypeId="1" AcceptedAnswerId="13"
  CreationDate="2010-09-01T19:34:48.000" Score="100" ViewCount="25786"
  Body="&lt;p&gt;A coworker of mine believes that &lt;em&gt;any&lt;/em&gt; use c
in-code comments&lt;/p&gt;&#xA;" OwnerUserId="6" LastEditorUserId="226"
  LastEditDate="2011-11-25T22:32:41.300"
  LastActivityDate="2012-11-27T19:29:27.740" Title="&quot;Comments are a code
smell&quot;" Tags="&lt;java&gt;&lt;python&gt;" AnswerCount="34"
  CommentCount="10" FavoriteCount="49" ClosedDate="2012-11-27T20:11:51.580"
  CommunityOwnedDate="2011-01-31T09:04:54.130" />
```

Characteristics of Data Set

Total number of ...

Posts	38,315
Tags	1,614

Tags per post

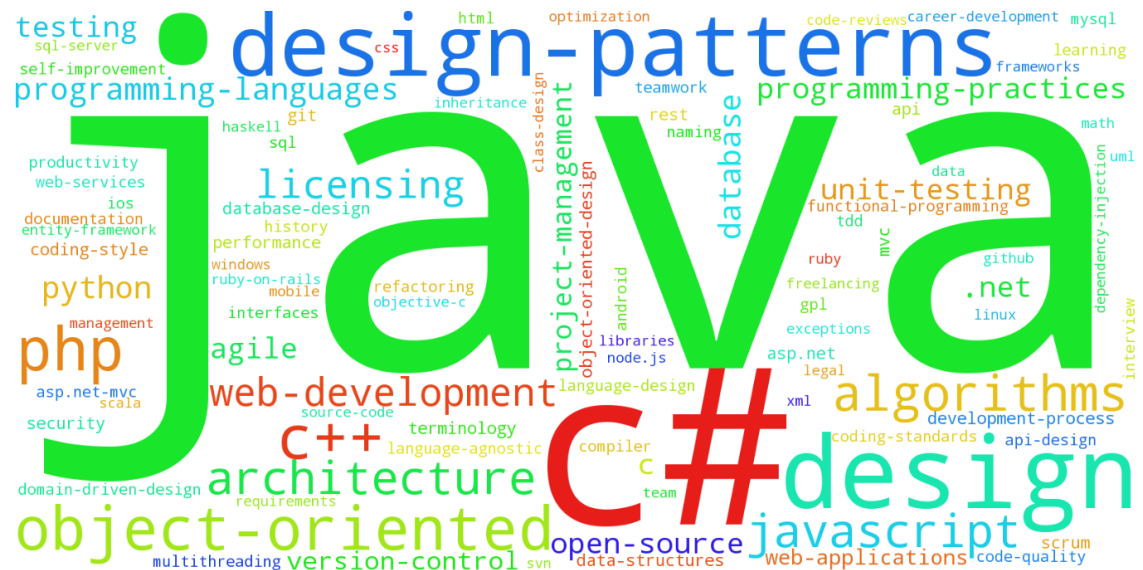
Min	1
Max	5
Average	2.68

Post length (words)

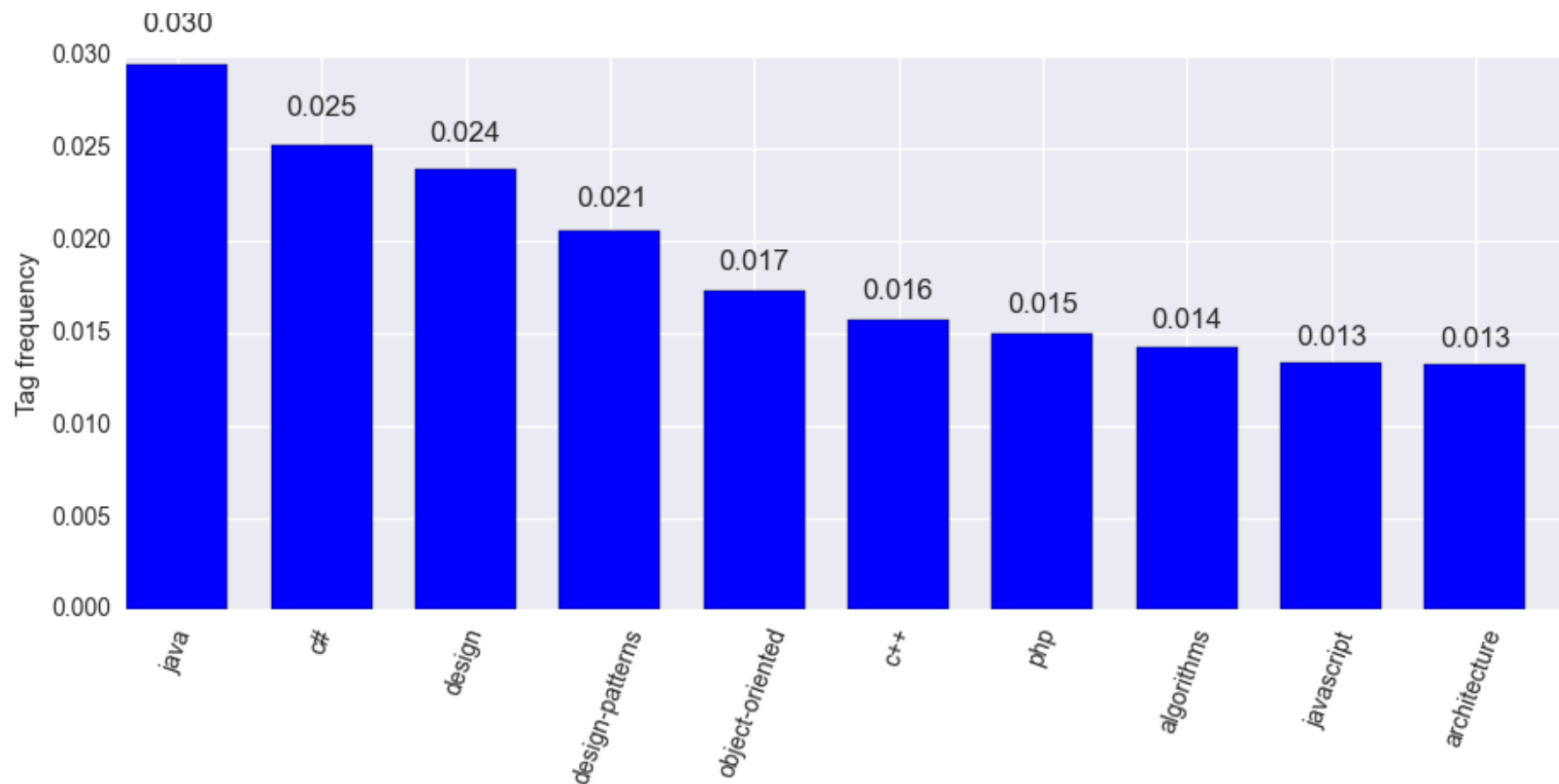
Min	6
Max	3,348
Average	178.65

Top 10 Tags

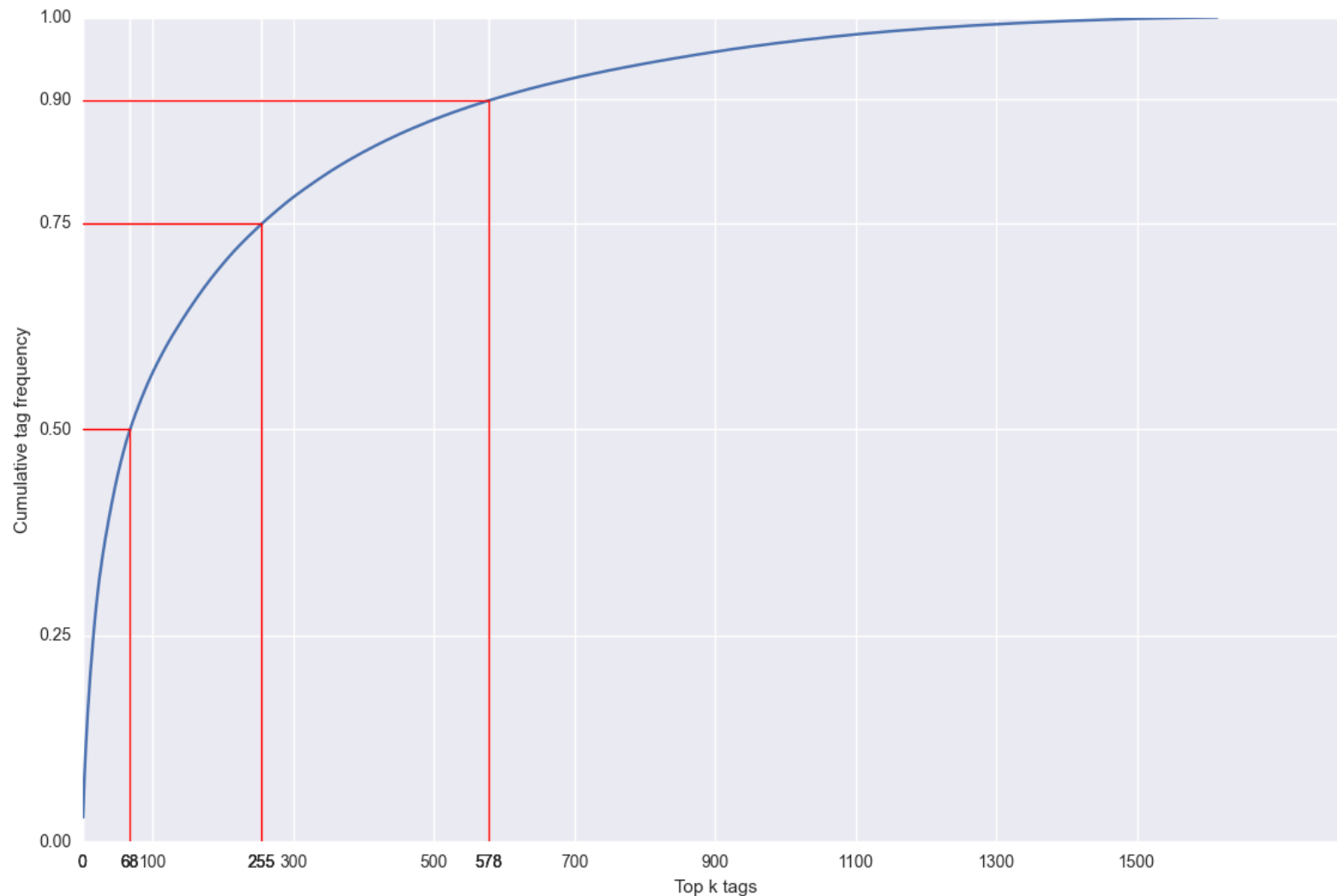
#	Tags
1	java
2	c#
3	design
4	design-patterns
5	object-oriented
6	c++
7	php
8	algorithms
9	javascript
10	architecture



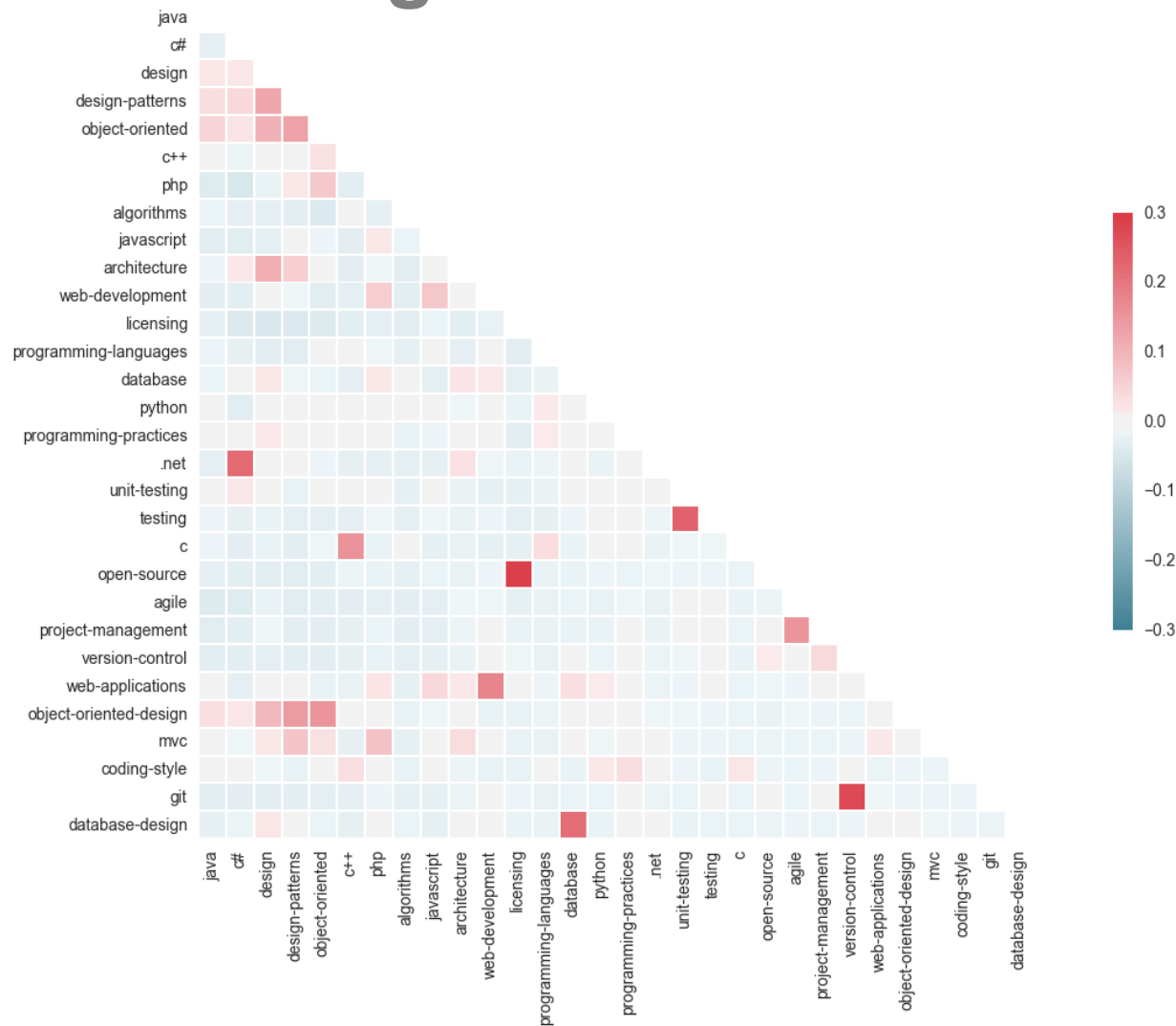
Top 10 Tags



Cumulative Tag Frequency



Tag Correlation



Outlook

Preprocessing

- Parse XML from Stack Exchange Data Dump (title, body, tags)

- Content
 - Strip html code from body
 - Remove special characters (keep care e.g. “C++” != “C”, “C#” != “C”)
 - Tokenize title + body
 - Stop word removal
 - Stemming
 - Lemmatization (e.g. “Windows” (OS) does not mean “window”)
 - POS-tagging
- Tags
 - Filter “rare”/“unique” tags → reduce dimensionality
 - Structure related tags (synonyms from Stack Exchange Data Explorer)

Outlook

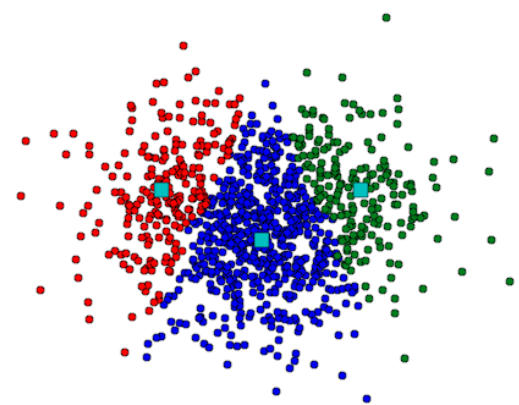
Supervised learning

- Separate data into test & training data
- Fit a supervised learning model
 - Naïve Bayes
 - Support Vector Machines (SVM)
 - k-Nearest Neighbors
- Evaluate best model
 - Precision
 - Recall
 - F1 measure
 - Accuracy

Outlook

Unsupervised learning

- Fit a model
 - k-means clustering
 - Hierarchical Agglomerative Clustering (HAC)
- Calculate TF-IDF (weighting)
- Evaluate best model
 - Precision
 - Recall
 - F1 measure
 - Rand Index



Thank you for your attention.

Attachment 1: Tag Correlation (zoomed)

