Data
6,034,196 questions
5,669,244 questions containing top 1000 tags (93%)
17,409,995 tag occurrences
12,454,534 top 1000 tags occurrences (71.5%)
2.8852 tag/question

1percent, 0_1000, C=10, max_features=5000 (L1)
4296.0 6379.0 21029.0
Precision: 67.345979 %
Recall: 20.428931 %
Mean F1: 0.244702

1percent, 0_1000, C=16, max_features=5000 (count_vectorizer)
6646.0 28940.0 21029.0
Precision: 22.964755 %
Recall: 31.603975 %
Mean F1: 0.258721

1percent, 0_1000, C=16, max_features=5000 (binary)
6345.0 13525.0 21029.0
Precision: 46.913124 %
Recall: 30.172619 %
Mean F1: 0.309113

1percent, 0_1000, C=10, max_features=5000 (binary)
6238.0 12315.0 21029.0
Precision: 50.653674 %
Recall: 29.663798 %
Mean F1: 0.311327

1percent, 0_1000, C=16, max_features=5000 (title only)
6331.0 13916.0 21029.0
Precision: 45.494395 %
Recall: 30.106044 %
Mean F1: 0.309396

1percent, 0_1000, C=16, max_features=10,000
6566.0 13736.0 21029.0
Precision: 47.801398 %
Recall: 31.223548 %

Mean F1: 0.3224931

1percent, 0_1000, C=10, max_features=10,000
6500.0 12654.0 21029.0
Precision: 51.367157 %
Recall: 30.909696 %
Mean F1: 0.325894

1percent, 0_1000, C=256
6294.0 18267.0 21029.0
Precision: 34.455576 %
Recall: 29.930097 %
Mean F1: 0.283419

1percent, 0_1000, C=128
6514.0 17603.0 21029.0
Precision: 37.005056 %
Recall: 30.976271 %
Mean F1: 0.294581

1percent, 0_1000, C=64
6516.0 16535.0 21029.0
Precision: 39.407318 %
Recall: 30.985782 %
Mean F1: 0.299992

1percent, 0_1000, C=32

1percent, 0_1000, C=16
6074.0 12714.0 21029.0
Precision: 47.774107 %
Recall: 28.883922 %
Mean F1: 0.300326

1percent, 0_1000, C=8
6182.0 11761.0 21029.0
Precision: 52.563558 %
Recall: 29.397499 %
Mean F1: 0.311546

1percent, 0_1000, C=4

Precision: 58.993162 %
Recall: 28.308526 %
Mean F1: 0.310130

1percent, 0_1000, C=2
5636.0 8708.0 21029.0
Precision: 64.722095 %
Recall: 26.801084 %
Mean F1: 0.303441

1percent, 0_1000, C=1
5039.0 7320.0 21029.0
Precision: 68.838798 %
Recall: 23.962148 %
Mean F1: 0.281697

5percent, 0_1000
(Real test)
Number of questions: 2013338.0
Number of tags predicted: 1575319.0
Number of question without tags: 864184.0
Tag per question: 0.782441398315
0.26707

5percent, 0_1000
(Real test, binary, C=16, max_features=10000)
Number of questions: 2013338.0
Number of tags predicted: 2951532.0
Number of question without tags: 386742.0
Tag per question: 1.46598931724
0.34670

- Use binary?
    - about the same
- Frequent itemsets => expand tag sets?
- Feature selection
- Try title
    - almost as good
- Word count instead of tf-idf
    - worse
- L1 norm
    - worse
- 10,000 vocab
    - better
- include title
- try RBF