1. Details for forum data: Number of tokens in each vertical [O]

3 forums for which we have the most posts in each vertical [O]

|  |  |  |
| --- | --- | --- |
| **Forum** | **# lines** | **# tokens** |
| **Automotive** | 29285064 = 29M | 2356778605 = 2B |
| **Collectables** | 19064351 = 19M | 379432068 = 380M |
| **Home garden** | 901958 = 902K | 51447856 = 5M |
| **Online health wellness** | 3752912 = 3.7M | 294419672 = 294M |
| **Outdoors** | 6858350 = 6.9M | 626109660 = 626M |
| **PE** | 616 | 39643 = 39K |
| **Pets** | 3570695 = 3.6M | 270395330 = 270M |
| **Power Sports** | 7308281 = 7.3M | 415526576 = 416M |
| **Sports** | 7180706 = 7.2M | 329590747 = 330M |
| **Technology** | 2073377 = 2M | 308679378 = 309M |

1. Details for parameter exploration for word2vec on analogy task.

Skipgram

Dimensionality: 200

Sampling: 0

Window size: 5

These parameters are mostly default values.

1. Add accuracy to Table 5 [O]

-Table5 updated.

1. Figure out P, R, and F for MFC baseline for Table 5. [O]

-The baseline classifier's default setting was set to “stratified” to generate predictions (based on the training set's class distribution). After changing this setting to “most\_frequent”, observed P,R and F scores expectedly. Changes updated to table 5.