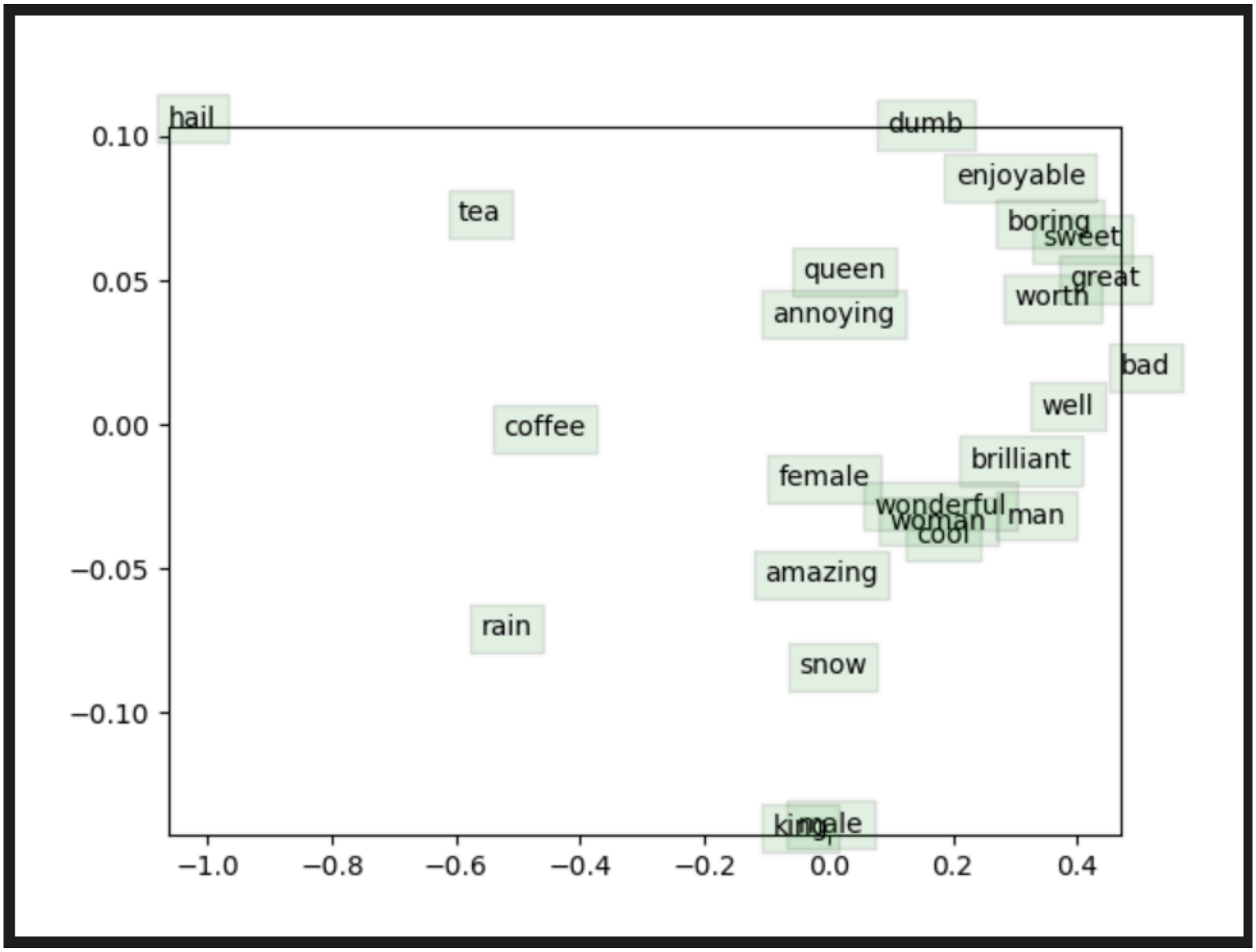
Q2



I notice that women are really cool, amazing, wonderful (i totally agree)
The trends and clusters are more grammatic than semantic. “Boring” and “enjoyable” serve very similar grammatic purposes but completely different semantic meanings, so i’d rather want them to be separated.

3a.

$\#(c, o)$ is the number of co-occurences of c and o.

for all pair of c, o that apear in the same window, we have another term of $p_\theta(o|c)$ in $\mathcal{L}$.

we have:

$$\mathcal{L}(\theta) = \prod_{i=1}^{T}\prod_{j=1}^{T} p_\theta(w_i|w_j)^{\#(c,o)}$$

$$J(\theta) = \sum_i \sum_j \#(w_i, w_j) log(p_\theta(w_i|w_j))$$

distinction:

$$\max_\theta J(\theta) = \max_\theta \sum_i \sum_j \#(w_i, w_j) log(p_\theta(w_i|w_j)) = \sum_j \max_\theta \sum_i \#(w_i, w_j) log(p_\theta(w_i|w_j))$$

that is because for all j, $p_\theta(\cdot \mid w_j)$ is an independent probability function (that is, for all choice of functions for $p_\theta(\cdot \mid w_{\tilde{j}\neq j})$, we can choose $p_\theta(\cdot \mid w_j)$ to be any probability function).

now let us take a specific c $(=w_j)$.

$$J_c(\theta) = \sum_i \#(w_i, c) log(p_\theta(w_i|c))$$

$$\nabla J_c(\theta) = \sum_i^T \#(w_i, c) \frac{\nabla p_\theta(w_i|c)}{p_\theta(w_i|c)}$$

we also have (for free!):

$$\sum_i^T p_\theta(w_i|c) = 1$$

so let us take $g_c(\theta) := \sum_i^T p_\theta(w_i|c) - 1$ and use lagrange multipliers:

$$\nabla g_c(\overrightarrow{p_\theta(\cdot|c)}) = (1, 1, ..., 1)$$

$$\nabla J_c\left(\overrightarrow{p_\theta(\cdot|c)}\right) - \lambda \nabla g_c(\overrightarrow{p_\theta(\cdot|c)}) = \left(\frac{\#(w_i, c)}{p_\theta(w_i|c)} - \lambda\right)_i$$

But we also know that θ^* is argmax globaly, so $\overrightarrow{p_{\theta^*}(\cdot|c)}$ is a global maximum of J_c, that is, $\nabla J_c\left(\overrightarrow{p_\theta(\cdot|c)}\right) = 0$.

moreover, $\nabla g_c(\overrightarrow{p_\theta(\cdot|c)}) = \nabla 0 = 0$.

so we have:

$$0 = \left(\frac{\#(w_i, c)}{p_\theta(w_i|c)} - \lambda\right)_i$$

and in particular, for all i:

$$0 = \frac{\#(w_i, c)}{p_\theta(w_i|c)} - \lambda$$

$$p_\theta(w_i|c) = \frac{1}{\lambda} \cdot \#(w_i, c)$$

and since $\sum_i^T p_\theta(w_i|c) = 1$:

$$1 = \sum_i^T p_\theta(w_i|c) = \frac{1}{\lambda} \cdot \sum_i^T \#(w_i, c)$$

$$\lambda = \sum_i^T \#(w_i, c)$$

remembering that $p_\theta(w_i|c) = \frac{1}{\lambda} \cdot \#(w_i, c)$, we get that

$$p_\theta(w_i|c) = \frac{1}{\lambda} \cdot \#(w_i, c) = \frac{\#(w_i, c)}{\sum_i^T \#(w_i, c)}$$

3b.

$$P(o|c) = \frac{e^{v_c u_o}}{\sum e^{v_c u_k}} = \frac{e^{u_o}}{\sum e^{u_k}}$$

so for all o, c, d

$$P(o|c) = P(o|d)$$

for example let us take $\{"ad", "ab", "cb"\}$
in this case we have

$$P(a|d) = 1 \neq 0.5 = P(a|b)$$

so we can't implement the ampirical probability using the mentioned model, and we seen that the ampirical distribution is the optimal (Most Likely).

4a.

The output of relu is a non negative vector, so the dot product results in a non negative number, so the sigmoid results in a value bigger than 0.5, so the model always predicts "true", so the ratio between successes and failures is 1:2.

4b.

get rid of activation, or use leaky-relu, or some non-non-negative activation.

4c.

Accuracy is bad since the dataset is inbalanced. recall by itself wont be good enough, since one can allow himself to guess positively all the time. precision alone is again not good enough, because it tolerates many false negatives. AUC methods wont belong here because it is not feasable to measure the tradeoff curves. so we are left with confusion matrix.

Question 6

1.
    a.
        i. C_v: Given a topic modeling, we take the N most common words in each topic and check how semanticly connected are they by checking if they appear in similar context (it's the same idea as behind word2vec). For this it uses NPMI. NPMI Measures how much are two words occurring together. I seems to me similar to the notion of covariance.
        ii. MACC: same idea as C_v, it checks the similarity between the top N most common words in each topic, but now using the cosine of two words's embedings (with a pretrained embedder)

Benchmarks

Add a Result

These leaderboards are used to track progress in Topic Models

Trend	Dataset	Best Model	Paper	Code	Compare
	AG News	DeTiME			See all
	20NewsGroups	vONTSS			See all
	20 Newsgroups	Bayesian SMM			See all
	Arxiv HEP-TH citation graph	JoSH			See all
	NYT	JoSH			See all
	AgNews	vONTSS			See all

    b.
        AG News, 20NewsGroups, NYT
2. .
3. c_v: 0.41, c_npmi: -0.17
4. I think that the words that are not the most common in each topic are also important. Another thing that a human reader could distinguish between two words that have different semantics but do appear a lot together, like “good” and “job”, or “my” “god”, but these measures doesnt.