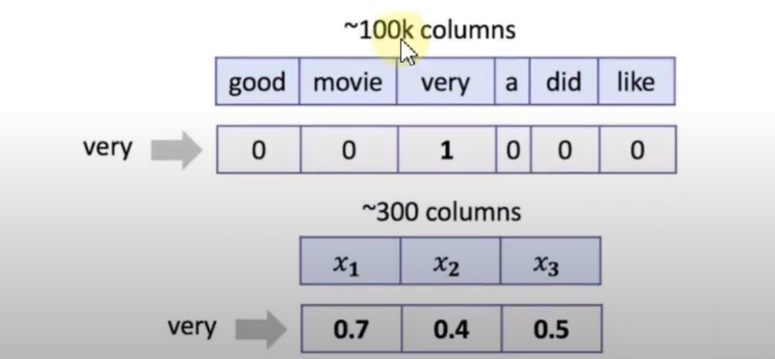
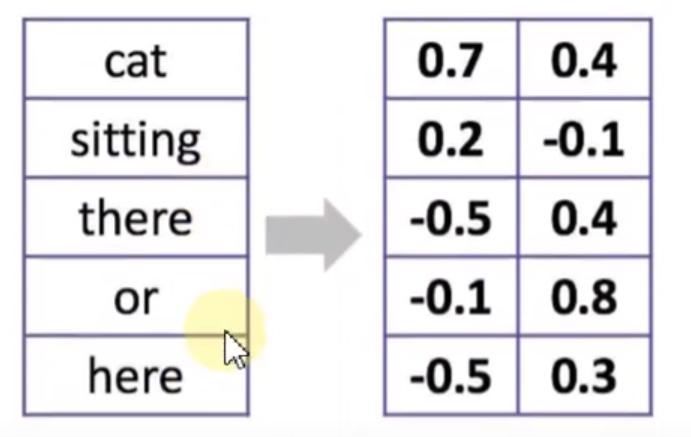
Perguntas

1. **Let’s recall how we treated words as one-hot sparse vectors in BOW and dense embeddings in neural networks:**

* Linear Model on top of a sum of neural representations can work faster than on top of BOW. (Correct – We only need to train 300 parameters here. Don’t forget no normalize these features row-wise)
* For both word representations we can take a sum of vectors corresponding to tokens of anu text to obtain good features for this text for further usage in linear model. (Correct)
* For both word representations we can take a **weighted** sum of vectors corresponding to tokens of anu text to obtain good features for this text for further usage in linear model. The weight for any token can be a TD-IDF ??? for that token (Correct – For BOW we effectively get bag of TF-IDF values, where TF is a binary variable.)
* You can replace **word2vec** embeddings with any **random** vectors to get a good features descriptor as a sum of vectors corresponding to all text token. (Incorrect)

|  |  |
| --- | --- |
| 1 | 0 |
| 0 | 1 |

|  |
| --- |
| 0.6 |
| 0.6 |
| 0.3 |
| 0.1 |

1. L**et’s recall 1D convolutions for words. What is the result of 1D convolution + maximum pooling over time for the following kernel without padding:**

=

X

Answer: **0.6**

1. **Let’s recall 1D convolutions for characters. Choose correct statements:**

* 1D convolutions work better than BOW for huge datasets (Correct)
* One 1D convolution layer for spotting character 3-grams is enough for solving a practical task. (Incorrect)
* 1D convolutions for characters consume one-hot encoded vectors for characters. (Correct)

1. **Choose true statements about text tokens:**

* A model without stemming/lemmatization can be the best. (Correct – Word2vec embeddings, for instance, are trained on raw tokens)
* Lemmatization needs more storage than stemming to work. (Correct – You must store information about all possible word forms in the vocabulary)
* Stemming can be done with heuristic rules (Correct)
* Lemmatization is always better than stemming (Incorrect)

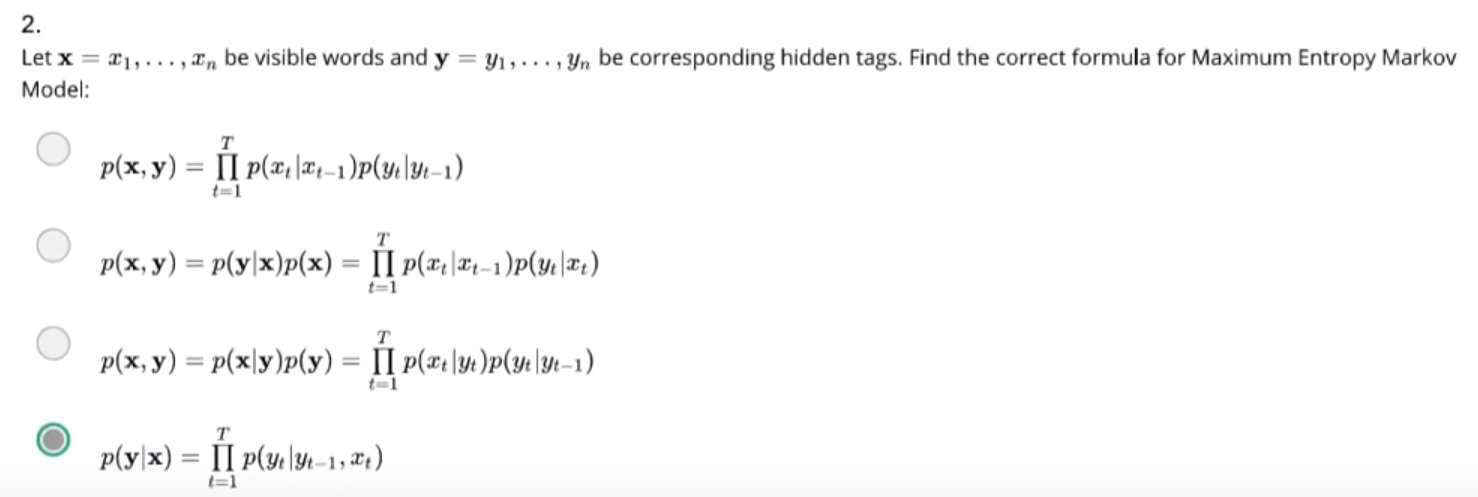
1. **Choose correct statements about BOW (or n-grams) features:**

* We prefer **sparse** storage formats to BOW features (Correct – We have a lot of zeros in these features, that’s why we can store them efficiently in sparse formats)
* You get the same vectorization result for any words permutation in your text (Incorrect)
* Classical **BOW** **vectorizer** (object that does vectorization) needs an amount of RAM at least proportional to T, which is the number of unique tokens in the dataset. (Correct – You must store a hasp map {token: index} to be able to vectorize new texts)
* **Hashing vectorizer** need an amount of RAM proportional to vocabulary size to operate (Incorrect – It only need to store a hash function)
* For **BOW features** you need an amount of RAM at least proportional to N x T, where N is the number of documents, T is the number of unique tokens in the dataset (Incorrect – You can and **should** store these features in sparse matrix)

1. **Find one incorrect statement below:**

* End-of-sentence tokens are necessary for modelling probabilities of sentences of different lengths
* If a test corpus does not have OOV words, smoothing is not needed (THIS ONE! Even though the probabilities will not be equal to 0, they will still poorly evaluated for poor terms)
* The smaller holdout perplexity is – the better the model.
* Trigram language models can have a larger perplexity than bigram language models.
* N-gram language models cannot capture distant contexts.

1. **Which of these models are discriminative, i.e., which of them model the distribution p(y|x)?**

* HMM
* CRF (Correct)
* MEMM (Correct)

1. **How many parameters does PLSA topic model have?**

* |T| x |W| + |T|\*|D|

T – Number of Topics | W – Vocabulary size | D – Number of documents

1. **Which assumptions are made in PLSA topic model?**

* Conditional independence: p(w | t,d) = p(w | t)
* BOW assumption

1. **Choose correct statements about SVD:**

* Squares of singular values of a matrix X are eigenvalues of XTX (or XXT) (Correct)
* Singular values of a rectangular matrix are its eigenvalues (Incorrect)
* Truncated SVD is the best rank $k$ approximation of the original matrix in term of Frobenius norm (Correct)
* Singular values can be negative (Incorrect)
* Any rectangular matrix with real entries has a SVD (Correct)
* SVD is not unique (for example, the zero matrix can be decomposed in infinitely many ways) (Correct)

1. **How are word embeddings usually evaluated (qualitatively or quantitatively)?**

* By the interpretability of the components of the vectors (Correct)
* By the number of positive components of word vectors (Incorrect)
* By Spearman’s correlation (or similar rank correlation measure) with human judgements on word similarity task (Correct)
* By the accuracy of analogy prediction (using some pre-defined dataset of 4-word analogies) (Correct)
* By comparing maximal lengths of word vectors (the more is the length, the better is the model) (Incorrect)

1. **Choose the correct statements:**

* For word similarity tasks, count-based methods perform on par with predictive models (Correct)
* Word2vec works fine for word analogies, but there are many concerns with word similarities (Incorrect)
* Representation of word or character n-grams may improve the quality of the model (Correct)
* Skip-gram negative sampling (SGNS) model is too hard to train, and it is often approximated with softmax (Incorrect)

1. **Which techniques would help if the data has rich morphology, informal spelling, and other** sources of OOV tokens?

* Hierarchical softmax (Incorrect)
* Copy mechanism (Correct)
* Negative sampling (Incorrect)
* Sub-word modeling (Correct)
* Byte-pair encoding (Correct)

1. **Find correct statements below:**

* Recent machine translation systems provide equally good quality for all language pairs (Incorrect)
* Machine Translation area was developing with gradual advances each year (Incorrect)
* “Interlingual” level of transfer provides the best accuracy in statistical machine translation systems (Incorrect)
* Neural Machine Translation is able to produce translations for language pairs that have never been observed in train (Correct)
* Evaluation in Machine Translation is hard, mostly because of many variations in translations (Correct)

1. **Let us say we are building a translation system from Greek (g) to Bulgarian (b). Which of the following statements are correct?**

* Language model here is complicated because different word alignments are possible (Incorrect)
* The noisy channel concept here corresponds to conditional distribution p(g|b) (Correct)
* We will need to build a translation model p(b|g) (Incorrect)
* We will need to build a translation model p(b) (Correct)

1. **What is considered a part of NLU?**

* Intent Classifier (Correct)
* Slot tagger (Correct)
* State tracker (Incorrect)

1. **What metrics do we use for NLU evaluation?**

* Number of turns in the dialog (Incorrect)
* Slots F1 (Correct)
* Task success rate (Incorrect)
* Intent accuracy (Correct)

1. **Choose correct statements about NLU:**

* Joint NLU model can’t product predictions faster than two separate models combined (one for intent classification and another for slot tagging) (Incorrect)
* Training a joint NLU model helps intent classifier and slot tagger. (Correct)
* You can use Convolutional Networks for slot tagging (Correct)
* You can use 1D Convolutions for intent classification (Correct)

1. **Choose correct statement about dialog context:**

* We can use memory networks to deal with context (Correct)
* We can add a simple feature like “previous utterance intent” as a categorical feature to NLU to start taking into account the context of the dialog (Correct)
* We need dialog context in single-turn dialogs (Incorrect)