## **1.**English

from scratch从头开始 manually人工地 modality形式

compress压缩 mitigate the sparsity减轻稀疏性 sentiment情感

semantics语义学 invariant不变的 scalar 标量

## **2.**Intro

1.NLP applications: Machine Translation, Sentiment Analysis, Speech Synthesis(语音合成)

pre-traing : generic skills

fine-tuning : specific skills

## **3.**Regular expression:

[Aa] : A or a

[A-Z]:A to Z

Negation :

[^Ss] : neither ‘S’ nor ‘s’

[^A-Z] : not upper letter

[^e^] : neither ‘e’ nor ‘^’

a^b : a^b

|a|b|c : [abc]

colou?r : color colour

o\*h! : 0 or more of previous char (ooo)h!

o+h! : 1 or more o(ooo)h!

. : any char or num

^ means start , $ means end :

^[^A-Za-z] : 1Hello

\.$ : the end. ‘\’ escape its special meaning 转义

.$ : the end!

## **4.**Text Normalization

Word Tokenization : space-based , without space

Word Normalization :

IR : reduce all letters to lower case

possible exception : upper case in mid-sentence, like USA

Stemming(词干提取) chops off affixes(词缀) crudely, while lemmatization(词形还原) reduces words to their dictionary headword form, and they may not yield identical results

Porter Stemmer(波特词干提取法)

Sentence Segmentation : tokenize first , an abbreviation dictionary(缩写词典) can help

## **5.**N-grams

1.language has long-distance dependencies,N-grams is insufficient

2.Maximum Likelihood Estimate(MLE):

3.intrinsic evaluation: perplexity, inverse probability of the test set

4.Zeros:One of the generalization,things that dont occur in the train set,but in the test set

5.Add-1 smoothing: add 1 to all the counts,P(Wi|Wi-1)=, but it can overly flatten the probability distribution

6.Backoff and interpolation: Backoff(回退) enables the model to rely on lower-order(低阶) N-grams when there is inadequate data to estimate the probability of a higher-order N-gram. Interpolation(插值) combines probabilities from different N-gram orders to estimate the probability of a sequence of words, λ.

## **6.**Vector Sematics

1.generalize similar but unseen words

2.Cosine Similarity

3.Bag of Words: position doesn’t matter, only count word frequency

4.Sparse Embeddings: tf-idf , tf(t,d)= , =, is the number of documents t occurs in.

When computational resources are limited, it might be preferred

5.Dense Embeddings: Word2Vec , skim-gram model , context relation

## **7.**Naive Bayes and Logistic Regression

1.generative classifiers: E.g 2 models , 1 picture, which one fits better

discriminative(判别) classifier: distinguish dogs from cats

2.Naive Bayes: generative linear classifier, use add-1 smooth

ignore the unknown words, use all words and dont use stopword lists

is the count of Cj,

is the count of all words of c, |V| is the count of vocabulary

=

3.Logistic Regression: discriminative linear classifier

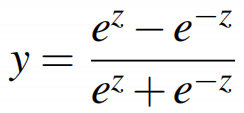


4.In text classification, the potential drawbacks of using a very large number of features: overfitting, underfitting, slower training speed

5.factors affect the performance of a text classification model: Quality of training data , Choice of classification algorithm , Size of the vocabulary

## **8.**Neural Network

1.Activation Functions: Sigmoid:  derivative=y(1-y)

tanh:  ReLu: 

2.steps in training a neural network for NLP tasks: Parameter Initialization , Forward Pass Backward Pass

## **9.** Possible Answer Questions

1.Briefly explain the difference between unigram, bigram, and trigram models.

Unigram models consider the probability of a single word. Bigram models condition on the previous word to predict the next. Trigram models take into account the two previous words for prediction. They differ in the amount of context they use to estimate word probabilities.

2.What is the purpose of the Laplace (add - 1) smoothing in Naive Bayes classification?

The purpose is to handle the problem of zero probabilities. By adding 1 to the count of each word in the training data, it ensures that no word has a zero probability, allowing the model to make predictions for unseen words.

3.How does Word2Vec learn word embeddings?

Word2Vec learns word embeddings by predicting whether a word is likely to appear nearby a target word. It trains a classifier on a binary prediction task and uses the learned classifier weights as the word embeddings.

4.Explain the concept of perplexity in the context of language models.

Perplexity measures how well a language model predicts the next word in a sequence. It is calculated as the inverse probability of a test set, normalized by the number of words. A lower perplexity indicates a better model.

5.What is the role of the bias term in a neural unit?

The bias term allows the neural unit to shift the activation function. It provides an additional degree of freedom in the model, enabling it to better fit the data by adjusting the output.

6.What is the difference between generative and discriminative classifiers?

Generative classifiers model how data is produced for each class, while discriminative classifiers focus on finding the optimal decision boundary to distinguish between classes. Generative classifiers learn the joint probability of data and classes, and discriminative classifiers learn the conditional probability of classes given the data.

7.What is the significance of cosine similarity in vector semantics?

Cosine similarity measures the similarity between two word vectors. It is calculated as the normalized dot product of the vectors and ranges from 0 to 1. Higher cosine similarity indicates more similar word meanings in the vector space.

8.How does the chain rule of probability apply in N - gram models?

The chain rule is used to calculate the joint probability of a sequence of words in N - gram models. It breaks down the probability of a sentence into the product of conditional probabilities of each word given the previous N - 1 words, but in N - gram models, we often make simplifying assumptions based on the Markov Assumption.

9.Why is text normalization important in NLP?

Text normalization is important as it prepares the text data for further processing. It helps in tasks like tokenization (breaking text into words), normalizing word formats (e.g., case folding, lemmatization), and sentence segmentation, which are essential for accurate analysis and model training.

10.Briefly describe the process of training a logistic regression model for text classification.

First, represent the text as feature vectors. Then, initialize the weights and bias. Calculate the output using the weighted sum and logistic function. Compute the cross - entropy loss and use gradient descent to update the weights based on the derivative of the loss function with respect to the weights. Repeat until convergence.