ELL884: Quiz 1

Answer the questions in the space provided. No partial marks will be awarded so write in detail.

Name: ____________________ Entry Number: ____________________

Total Marks: 50 **Time: 60 minutes** **Mon, Feb 17, 2025**

Question 1: RegEx

1. Suppose you have to create your own command line argument parser. You want to extract key-value pairs with optional fields. Write a single regex that captures username, action, and success if present, each as a group. Examples are given as, **(2 marks)**

```
user=Alpha action=login success=1
user=Beta action=download
user=Charlie success=0
user=Delta success=0 action=upload
```

2. You want to accept strings like `2.3e4`, `-4.56E+2`, `3.14E-10`, `3e3` but reject invalid forms such as `2.3e`, `.e4`, or `3.14.15e2`. Write regex to capture this. **(1 mark)**

3. Construct a regex pattern that correctly matches all valid English plural nouns while avoiding false positives like *is*, and *has*. Explore all edge cases and justify your choices. **(2 marks)**

Question 2: N-grams

1. Given a vocabulary size V and a corpus of size N, derive the number of parameters required to train an n-gram model. **(2 marks)**

2. Suppose you have two corpora: - **Corpus A** (in-domain, small) - **Corpus B** (out-of-domain, large). You suspect that your test data is more similar to Corpus A, but you do not have enough training data in Corpus A to obtain reliable N-gram counts. Devise a mixture approach to combine bigram estimates from both corpora. How do you determine the mixture weight α using expectation maximization or cross-validation? **(3 marks)**

3. Assume you have a bigram model over a large vocabulary V. Instead of relying on traditional discrete smoothing (e.g., Laplace or Kneser–Ney), you decide to embed each word into a continuous space $\mathbb{R}^d$ and use a smooth function $f(\mathbf{e}_{w_{n-1}}, \mathbf{e}_{w_n})$ to parametrize $P(w_n \mid w_{n-1})$. Describe how you would guarantee that $P(\cdot \mid w_{n-1})$ forms a valid probability distribution for each w_{n-1} in the vocabulary. In other words, explain how to ensure non-negativity and that the probabilities sum to 1. **(5 marks)**

Question 3: Minimum Edit Distance

1. You are given two strings S_1 and S_2 of length m and n, respectively. The minimum edit distance between them is defined as the minimum number of insertions, deletions, and substitutions required to transform S_1 into S_2. Modify the standard $O(m \cdot n)$ space DP solution to an $O(\min(m, n))$ space-efficient approach. Give your answer as pseudocode. **(3 marks)**

2. Instead of comparing two linear strings, now you need to compute the edit distance between a string and a Directed Acyclic Graph (DAG). For instance, you need to find the closest match between a DNA sequence and a reference genome represented as a DAG. The DAG represents multiple possible sequences. Your aim is to find the minimum edit distance from a given string to any path in the DAG. Extend the standard DP approach to handle DAG structures. **(5 marks)**

Question 4: Word Representations

1. Given word embeddings $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_N$ in a high-dimensional space, define a cosine similarity metric and prove that it satisfies the triangle inequality. **(3 marks)**

2. In word2vec negative sampling, frequent words are sampled with probability proportional to $P(w)^{\frac{3}{4}}$ rather than $P(w)$. Prove that this sampling reduces the variance of gradient estimates. **(5 marks)**

3. FastText represents a word as a sum of its subword n-grams. If a word w has length l, derive the number of possible 3-gram subwords. Extend this to arbitrary n-grams. Also define the boundary condition. **(3 marks)**

4. You were given a Word2Vec model in 2014 which you kept on training over time. Now that it has been trained for over a decade, propose a method to quantify semantic drift over time. **(3 marks)**

5. Bias in embeddings is often measured by: $\cos(v_{\text{man}}, v_{\text{programmer}}) - \cos(v_{\text{woman}}, v_{\text{programmer}})$. How can debiasing methods (e.g., Orthogonal Projection) remove gender bias while preserving semantic similarity? **(3 marks)**

Question 5: HMMs

1. Traditional HMMs assume discrete observation distributions $P(o_t|q_t) = B(q_t, o_t)$, where $B(q_t, o_t)$ represents the emission probability of observation o_t from state q_t. However, in many real-world applications (e.g., speech recognition), observations are continuous values, which limits the application of HMMs. Therefore, modify the standard HMM to support continuous emissions using Gaussian Mixture Models. The Gaussian probability density function is given as,

$$\mathcal{N}(o_t|\mu, \Sigma) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(o_t - \mu)^T \Sigma^{-1}(o_t - \mu)\right),$$

where d is the observation dimensionality, μ is the mean vector, Σ is the covariance matrix. Given that $c_{q_t}^{(i)}$ is the mixing weight for the i-th Gaussian, satisfying $\sum_i c_{q_t}^{(i)} = 1$, compute the formulae of new emission probability and derive new formulations for the forward and Viterbi algorithms, with initialization, recursion, and termination conditions clearly stated. **(10 marks)**

Write complete and exact answers with correct notations, no marks will be provided even if your answer is "closely accurate".