Examen Parcial N° 2

Período habilitado para la entrega : desde el miércoles 7 de julio hasta el domingo 11 de julio.

Modalidad

- El examen tiene 4 secciones de ejercicios.
- De cada sección deberá elegir cuáles preguntas contestar. **Debe dejar 4 preguntas sin contestar en cada sección.** EN CASO DE CONTESTAR MÁS PREGUNTAS DE LAS SOLICITADAS, SIEMPRE SE DESCARTARÁN RESPUESTAS CORRECTAS.
- Nota relativa Q1, Q2, Q3, Q4 según puntajes del curso.
- Puede usar bibliografía de cualquier tipo: libros, papers, foros de internet, clases, etcétera.
- La presentación y evaluación será individual.
- Por favor respete la numeración de las preguntas.
- Cree un documento pdf con las preguntas que eligió y respóndalas a continuación.
- Cualquier consulta con respecto a la comprensión de los enunciados se debe consultar en el foro #consultas-parcial2.

Sección 1: RNN, GRU, LSTM

1-1. Los enfoques secuenciales unidireccionales se basan en el pasado para predecir el siguiente token. ¿Qué limitaciones tiene este enfoque a nivel conceptual? De ejemplos de oraciones o palabras.

1-2. Además de texto, ¿qué otros problemas pueden ser atacados RNN (modelos autoregresivos con variables ocultas)?

1-3. Dado un dataset de 100,000 palabras, ¿Cuál es la frecuencia **máxima** para un conjunto cualquiera de 4 palabras?

1-4. Si agarramos un texto generado por un modelo entrenado para modelado del lenguaje, y ordenamos todas las palabras que aparecen en él en función de la frecuencia de aparición. ¿Qué relación espera que exista entre la frecuencia de aparición de cada palabra y su posición en el ranking para que se comporte como un lenguaje natural?

1-5. Si queremos que un ejemplo de secuencia sea una oración completa, ¿qué tipo de problema introduce esto en el muestreo de minibatch? ¿Cómo podemos solucionar el problema?

1-6. Si usamos un RNN para predecir el siguiente carácter en una secuencia de texto, ¿cuántas neuronas tiene la capa de salida?

1-7. ¿Qué sucede con gradiente durante la aplicación de“backpropagation” de una secuencia muy larga? Justifique. ¿Qué soluciones se utilizan?

1-8. Compare el costo computacional de GRU, LSTM y RNN convencionales para una dimensión de variable oculta k. Preste especial atención al costo de entrenamiento e inferencia.

1-9. Dado que el candidato de celda de memoria (C tilda) usa la función tanh para tener un rango de valores entre -1 y 1, ¿por qué la variable oculta (H) necesita usar la función tanh nuevamente para asegurarse de que el rango de valores de salida esté entre -1 y 1?

1-10. Hay lenguajes, como el chino y el japonés, donde no hay delimitadores entre palabras, es decir, no existen los espacios. ¿La tokenización a nivel de palabra sigue siendo una buena idea para tales casos? ¿Por qué o por qué no?

Sección 2: Embeddings

2-1. ¿Qué ventajas y desventajas tiene “one-hot encoding” en NLP?

2-2. ¿Cuál es la relación entre el producto interno de dos embeddings de palabras y la similitud del coseno?

2-3. Una palabra puede tener diferentes palabras de contexto o palabras de sampleo negativo en diferentes épocas. ¿Cuáles son los beneficios de este tipo de formación?

2-4. Explique en qué se basan los algoritmos como Word2Vec para generar ejemplos etiquetados que sirvan para aprender las relaciones semánticas presentes entre las distintas palabras.

2-5. ¿Qué diferencias existen entre Skipgram y Continuous Bag of Words? ¿Qué ventajas y desventajas ofrece cada alternativa?

2-6. ¿En qué consiste el muestreo negativo (negative sampling)? ¿Qué ventajas ofrece su implementación con la función sigmoidea en lugar de con softmax?

2-7. ¿Cuáles son los parámetros aprendibles en el proceso de entrenamiento de Word2Vec? ¿Y los hiperparámetros?

2-8. En el proceso de entrenamiento de embeddings GloVe, ¿qué rol cumplen la razones de co-ocurrencia entre las diferentes palabras en el armado del dataset? ¿Por qué se dice que el cálculo de estas razones de co-ocurrencia mejora los embeddings resultantes para palabras menos frecuentes?

2-9. ¿Cómo se obtienen los embeddings finales de Word2Vec a partir de las matrices de embeddings de palabra central y de contexto? ¿Y en GloVe? ¿A qué se debe la diferencia?

2-10. ¿En qué se diferencian los embeddings fastText de los enfoques anteriores como GloVe o Word2Vec? ¿Qué ventaja tiene este añadido?

2-11. ¿Qué problemas tenía la propuesta original de FastText de representar una palabra como una colección de n-gramas?

2-12. ¿En qué consiste la generación de subpalabras a partir de la codificación por pares de bytes?¿Qué ventajas trae con respecto a la propuesta original de FastText

Sección 3: Problemas secuencia a secuencia. Atención, Transformers y BERT

3-1. Supongamos que usamos redes neuronales para implementar la arquitectura encoder-decoder. ¿El encoder y el decoder tienen que ser del mismo tipo de red neuronal?

3-2. Además de la traducción automática, ¿puede pensar en otra aplicación en la que se pueda aplicar la arquitectura del encoder-decoder?

3-3. La polisemia es común en los lenguajes naturales. ¿Qué tipo de arquitectura neuronal es preferida para atacar este problema a la hora de generar embeddings?

3-4. ¿Por qué en el modelado de lenguaje usamos máscaras tanto en el decoder como en la función de pérdida? ¿Cómo sería la performance si no enmascaráramos?

3-5. Si en el entrenamiento establecemos un ratio de forzamiento del maestro en 0, ¿Cómo afectará esto a la performance del modelo?

3-6. Si usamos Beam Search para predecir la siguiente palabra, ¿cómo afecta el tamaño del haz a la calidad de los resultados y la velocidad de predicción?

3-7. Cuando los queries y las keys tienen la misma dimensión, ¿la suma de vectores es un mejor diseño que el producto escalar para la función de puntuación? ¿Por qué o por qué no? Nota: Hablamos de SUMA DE VECTORES, no de atención aditiva.

3-8. Si en lugar de usar una codificación posicional fija, como en transformers, aprendiéramos un embedding de codificación posicional, ¿que usaría como entrada a esas capas de embeddings?

3-9. ¿Por qué BERT usa encoders de transformers para modelado de lenguaje en lugar de decoders?

3-10. ¿Cuáles pueden ser los desafíos para los transformadores si las secuencias de entrada son muy largas? ¿Por qué?

3-11. En igualdad de condiciones, ¿un modelo de lenguaje con máscaras requerirá más o menos épocas para converger que un modelo de lenguaje de izquierda a derecha? ¿Por qué?

3-12. ¿Qué características tiene BERT? ¿Qué ventajas recupera de GPT y de ELMo?

3-13. La pérdida en modelado del lenguaje enmascarado para BERT es significativamente mayor que la pérdida de predicción de la siguiente oración. ¿Por qué?

Sección 4: Sistemas de recomendación

4-1. Si bien hay muchas herramientas para sistemas de recomendación, gran parte de los esfuerzos se basan en analizar la matriz de interacción. ¿Qué datos hay en esta matriz? ¿Cómo planteamos el problema de separar información del usuario y de los artículos a ofrecer?

4-2. En sistemas de clasificación hablamos de 3 tipos de funciones de pérdida. Esta clasificación surgía del tipo de entrada para una función de pérdida. ¿Cuáles fueron estos 3 tipos de funciones de pérdida? En el curso trabajamos mayormente con entropía cruzada y mínimos cuadrados. Ambas pertenecen a un mismo tipo de funciones de pérdida de la clasificación anterior, ¿a cuál?

4-3. Defina con sus palabras datos implícitos y datos explícitos en el contexto de sistemas de recomendación. ¿Cuáles de estos datos hacen al problema de filtros **colaborativos**?

4-4. La publicidad en línea muchas veces permite guardar información de cómo el usuario interactúa con el anuncio. ¿Cuáles de los siguientes eventos de mouse son datos implícitos y cuáles explícitos? Nota: recuerde que en publicidad el click-through-rate es una métrica usada para analizar el comportamiento del modelo.

a. mouseleave
b. mouseenter
c. mouseover
d. click
e. mouseup
f. contextmenu

4-5. El dataset MovieLens tiene embeddings con información de las películas como parte de los datos disponibles. ¿Cómo llama MovieLens a este embedding? ¿Que diferencia hay entre este embedding y los usados en el contexto de sistemas de recomendación?

4-6. ¿Qué representa cada término en una máquina de factorización de orden 2 como las vistas en clase? ¿Qué desventajas tienen las máquinas de factorización de orden mayor a 2? ¿Qué desventajas tienen las máquinas de factorización de orden a 1?

4-7. ¿Qué diferencia existe entre una máquina de factorización, una máquina de factorización profunda?

4-8. Defina con sus palabras las siguientes nociones de estadística:

a. Precisión
b. Especificidad
c. Exhaustividad
d. Exactitud
e. Matriz de confusión

4-9. ¿Por qué es posible usar un autoencoder en los sistemas de recomendación?

4-10. ¿Por qué se dice que la pérdida de bisagra considera el sampleo negativo?

4-11. ¿Por qué se dice que la pérdida bayesiana para ránkings personalizados es "bayesiana”? ¿Con qué técnica para evitar overfitting está relacionada?

4-12. Típicamente, ¿se usan datos implícitos o explícitos para hacer sampleo negativo?