PRÁCTICA DE MODELOS DE LENGUAJE

TAREA 1: Comparación de los diferentes modelos de lenguaje según el valor de N

Orden del Modelo	Resultado
1	1169 sentences, 8372 words, 0 OOVs, 0 zeroprobs, logprob= -18957.95, ppl= 97.05049, ppl1= 183.8427
2	1169 sentences, 8372 words, 0 OOVs, 0 zeroprobs, logprob= -10016.61, ppl= 11.2163, ppl1= 15.71962
3	1169 sentences, 8372 words, 0 OOVs, 0 zeroprobs, logprob= -8446.004, ppl= 7.677726, ppl1= 10.20562
4	1169 sentences, 8372 words, 0 OOVs, 0 zeroprobs, logprob= -8195.209, ppl= 7.226808, ppl1= 9.525396
5	1169 sentences, 8372 words, 0 OOVs, 0 zeroprobs, logprob= -8220.561 ppl= 7.271161, ppl1= 9.592047

Tabla 1: *Resultados para los modelos de lenguaje para los n-gramas*

TAREA 2: Good-Turing, Witten-Bell, modified Kneser-Ney y unmodified Kneser-Ney

Órden Good-Turing	Resultado
3	1169 sentences, 8372 words, 0 OOVs, 0 zeroprobs, logprob= -8446.004, ppl= 7.677726, ppl1= 10.20562
4	1169 sentences, 8372 words, 0 OOVs, 0 zeroprobs, logprob= -8195.209, ppl= 7.226808, ppl1= 9.525396
Órden Witten-Bell	**Resultado**
3	1169 sentences, 8372 words, 0 OOVs, 0 zeroprobs, logprob= -8571.477, ppl= 7.913772, ppl1= 10.56396
4	1169 sentences, 8372 words, 0 OOVs, 0 zeroprobs, logprob= -8179.678, ppl= 7.199771, ppl1= 9.484794

Órden Unmodified Kneser-Ney	**Resultado**
3	1169 sentences, 8372 words, 0 OOVs, 0 zeroprobs, logprob= -8449.501, ppl= 7.684207, ppl1= 10.21544
4	1169 sentences, 8372 words, 0 OOVs, 0 zeroprobs, logprob= -8121.629, ppl= 7.099612, ppl1= 9.334569
Órden Modified Kneser-Ney	**Resultado**
3	1169 sentences, 8372 words, 0 OOVs, 0 zeroprobs, logprob= -8688.563, ppl= 8.14058, ppl1= 10.90968
4	1169 sentences, 8372 words, 0 OOVs, 0 zeroprobs, logprob= -8409.026 ppl= 7.609513 ppl1= 10.10235

Tabla 2: *Resultados para los modelos de lenguaje usando los métodos de suavizado Witten-Bell y Modified Kneser-Ney bajo esquema backoff*

Tarea 3: Witten-Bell y modified Kneser-Ney, pero esta vez bajo un esquema de interpolación en lugar de backoff.

Órden Witten-Bell	**Resultado interpolado**
3	1169 sentences, 8372 words, 0 OOVs, 0 zeroprobs, logprob= -8372.384, ppl= 7.542518, ppl1= 10.00106
4	1169 sentences, 8372 words, 0 OOVs, 0 zeroprobs, logprob= -7952.97, ppl= 6.816434, ppl1= 8.911454
Órden Modified Kneser-Ney	**Resultado interpolado**
3	1169 sentences, 8372 words, 0 OOVs, 0 zeroprobs, logprob= -8439.978, ppl= 7.666568, ppl1= 10.18872
4	1169 sentences, 8372 words, 0 OOVs, 0 zeroprobs, logprob= -8026.721 ppl= 6.938845, ppl1= 9.094062

Tabla 3: *Resultados para los modelos de lenguaje usando los métodos de suavizado Witten-Bell y Modified Kneser-Ney bajo esquema de interpolación*

Conclusiones corpus Dihana

Tras analizar los resultados de los modelos de lenguaje con diferentes órdenes de n-gramas (véase en la Tabla 1), se puede observar que al evaluar los cinco modelos de lenguaje con órdenes distintas, el modelo de 1-grama obtuvo el peor rendimiento, con un alto valor de perplejidad (ppl) de 97.05049 y logprob negativo de -18957.95. Esto indica que el modelo de 1-grama tiene dificultades para capturar la estructura y las dependencias en el lenguaje. El modelo de 2-grama también mostró un rendimiento bajo, aunque mejor que el 1-grama. Esto sugiere que considerar sólo las palabras o los bigramas es insuficiente para modelar adecuadamente el lenguaje. Después, a medida que se aumenta el orden de los n-gramas, se observa una mejora significativa en el rendimiento. El modelo de 3-grama logró reducir la perplejidad a 7.677726 y mejorar el logprob a -8446.004. El modelo de 4-grama continuó mejorando los resultados, con una perplejidad de 7.226808 y un logprob de -8195.209. Sin embargo, el modelo de 5-grama, a pesar de ser más complejo, no logró superar al modelo de 4-grama, con una perplejidad de 7.271161 y el logprob fue -8220.561, lo que sugiere que aumentar más el orden de los n-gramas no necesariamente produce mejoras en el rendimiento.

Después, analizando los resultados obtenidos de los modelos de lenguaje usando los métodos de suavizado Witten-Bell, Good Turing, Unmodified y Modified Kneser-Ney para los 3-gramas y 4-gramas (véase en la Tabla 2), se observa que de manera general los 4-gramas tienden a funcionar mejor que los modelos de trigramas en términos de perplejidad y logprob. En general, los modelos que usan métodos de suavizado mejoran los modelos que se han comentado anteriormente, especialmente destacan los modelos de lenguaje con suavizado Witten-Bell y Unmodified Kneaser-Ney, este último obteniendo unos resultados ligeramente mejores que el anterior.

Por último, comparando modelos Witten-Bell y Modified Kneser-Ney bajo el esquema de suavizado backoff (véase en la Tabla 3) mencionados anteriormente, con estos mismos modelos pero bajo el esquema de interpolación. Se aprecia una gran mejoría, sobre todo con los 4-grama los cuales obtienen un rendimiento mejor, logrando reducir el logprob a -7952.97 y la perplejidad (ppl) a 6.816434.

De este modo, en resumen el órden que obtiene mejores resultados es el 4-grama, de manera general. Además, los métodos de suavizado tanto bajo el esquema backoff o de interpolación, logran disminuir los valores de logprob y perplejidad. El modelo que mejor pudo capturar las características del lenguaje fue el modelo 4-grama con método de suavizado Witten-Bell bajo un esquema de interpolación.