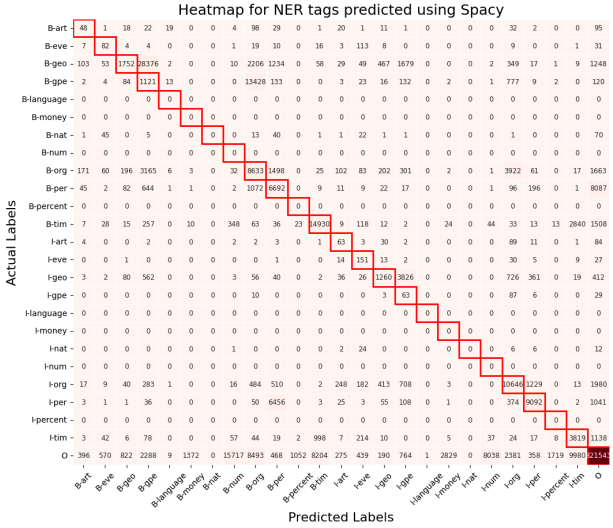
NLP Assignment 1

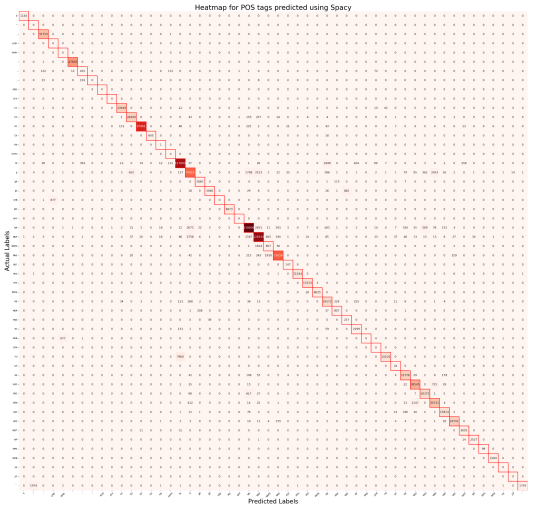
By Abhimanyu Sharma and Samveg Bhansali

Spacy Evaluation Plots

Heatmap for NER tags predicted using Spacy



Heatmap for POS tags predicted using Spacy



TagWise Evaluation Metrics for POS using Spacy

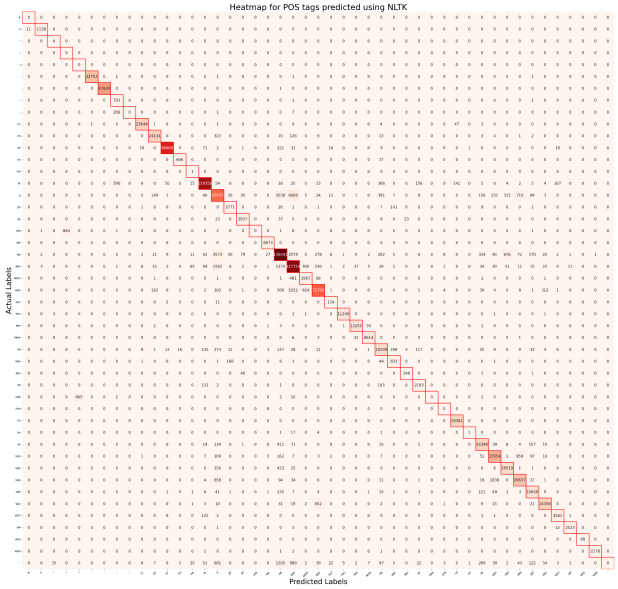
	TP	FP	TN	FN	Accuracy	Precision	Recall
$	1145	1	1043576	4	0.999995	0.999127	0.996519
,	32753	179	1011794	0	0.999829	0.994565	1.000000
.	47830	14	996881	1	0.999986	0.999707	0.999979
:	435	547	1043609	135	0.999347	0.442974	0.763158
;	0	0	1044512	214	0.999795	NaN	0.000000
CC	23649	189	1020831	57	0.999765	0.992071	0.997596
CD	24200	698	1019334	494	0.998859	0.971966	0.979995
DT	97961	112	946160	493	0.999421	0.998858	0.994993
EX	645	3	1044060	18	0.999980	0.995370	0.972851
FW	1	53	1044672	0	0.999949	0.018519	1.000000
IN	117693	8546	915384	3103	0.988850	0.932303	0.974312
JJ	70618	4945	961393	7770	0.987829	0.934558	0.900878
JJR	2840	244	1041515	127	0.999645	0.920882	0.957196
JJS	2068	50	1041642	966	0.999027	0.976393	0.681608
LRB	0	0	1044725	1	0.999999	NaN	0.000000
MD	6973	9	1037744	0	0.999991	0.998711	1.000000
NN	138808	4867	894064	6987	0.988653	0.966125	0.952077
NNP	126543	7317	906023	4843	0.988361	0.945338	0.963139
NNPS	857	2792	1039413	1664	0.995735	0.234859	0.339944
NNS	73059	729	968157	2781	0.996640	0.990120	0.963331
PDT	147	79	1044500	0	0.999924	0.650442	1.000000
POS	11244	4	1033466	12	0.999985	0.999644	0.998934
PRP	13313	57	1031351	5	0.999941	0.995737	0.999625
PRP$	8635	6	1036065	20	0.999975	0.999306	0.997689
RB	19173	3006	1021468	1079	0.996090	0.864466	0.946721
RBR	807	348	1043323	248	0.999430	0.698701	0.764929
RBS	257	892	1043538	39	0.999109	0.223673	0.868243
RP	2299	662	1041574	191	0.999184	0.776427	0.923293
RRB	0	0	1044725	1	0.999999	NaN	0.000000
TO	15101	51	1021614	7960	0.992332	0.996634	0.654828
UH	24	86	1044616	0	0.999918	0.218182	1.000000
VB	23776	696	1019819	435	0.998917	0.971559	0.982033
VBD	38545	1368	1003979	834	0.997892	0.965725	0.978821
VBG	18375	949	1024652	750	0.998374	0.950890	0.960784
VBN	30711	2870	1009528	1617	0.995705	0.914535	0.949981
VBP	15814	410	1028158	344	0.999278	0.974729	0.978710
VBZ	24742	377	1019389	218	0.999430	0.984991	0.991266
WDT	3655	285	1040743	43	0.999686	0.927665	0.988372
WP	2527	31	1042153	15	0.999956	0.987881	0.994099
WP$	99	0	1044627	0	1.000000	1.000000	1.000000
WRB	2184	1	1042541	0	0.999999	0.999542	1.000000
``	1745	2	1042973	6	0.999992	0.998855	0.996573

TagWise Evaluation Metrics for NER using Spacy

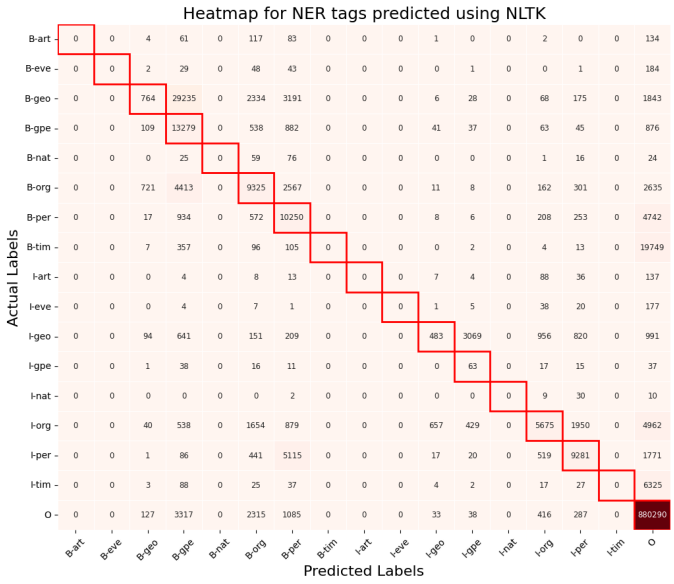
	TP	FP	TN	FN	Accuracy	Precision	Recall
B-art	48	762	1015993	331	0.998925	0.059259	0.126649
B-eve	82	817	1016010	225	0.998976	0.091212	0.267101
B-geo	1752	1349	978156	35877	0.963401	0.564979	0.046560
B-gpe	1121	35722	965560	14731	0.950397	0.030426	0.070717
B-nat	0	0	1016933	201	0.999802	NaN	0.000000
B-org	8633	26038	970997	11466	0.963128	0.248998	0.429524
B-per	6692	10477	989672	10293	0.979580	0.389772	0.393995
B-tim	14930	9320	987943	4941	0.985979	0.615670	0.751346
I-art	63	785	1016054	232	0.999000	0.074292	0.213559
I-eve	151	1309	1015572	102	0.998613	0.103425	0.596838
I-geo	1260	1453	1008270	6151	0.992524	0.464431	0.170018
I-gpe	63	7543	1009393	135	0.992451	0.008283	0.318182
I-nat	0	0	1017084	50	0.999951	NaN	0.000000
I-org	10646	8936	991434	6118	0.985200	0.543663	0.635051
I-per	9092	2291	997593	8158	0.989727	0.798735	0.527072
I-tim	3819	12892	997823	2600	0.984769	0.228532	0.594952
O	821543	17545	142418	35628	0.947723	0.979090	0.958435

NLTK Evaluation Plots

Heatmap for POS tags predicted using NLTK



Heatmap for NER tags predicted using NLTK



TagWise Evaluation Metrics for POS using NLTK

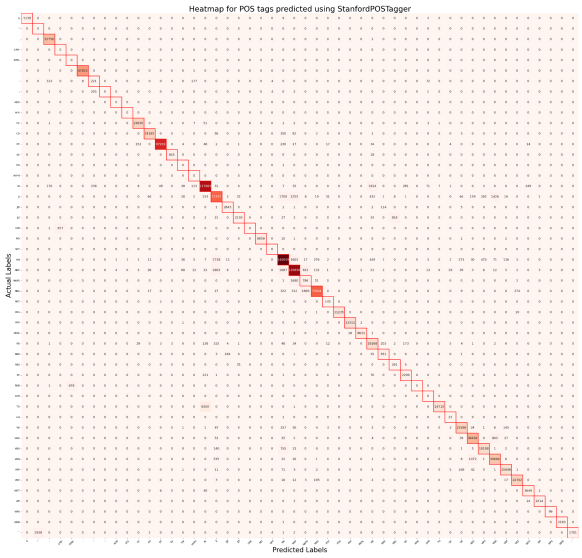
TagWise Evaluation Metrics for POS using NLTK							
	TP	FP	TN	FN	Accuracy	Precision	Recall
$	1138	6	1046036	0	0.999994	0.994755	1.000000
,	32753	1	1014422	4	0.999995	0.999969	0.999878
.	47830	0	999349	1	0.999999	1.000000	0.999979
:	793	800	1045585	2	0.999234	0.497803	0.997484
;	0	0	1046966	214	0.999796	NaN	0.000000
CC	23646	58	1023406	70	0.999878	0.997553	0.997048
CD	24144	379	1022106	551	0.999112	0.984545	0.977688
DT	98069	96	948630	385	0.999541	0.999022	0.996090
EX	646	20	1046497	17	0.999965	0.969970	0.974359
FW	1	113	1047066	0	0.999892	0.008772	1.000000
IN	119350	694	925490	1646	0.997765	0.994219	0.986396
JJ	67473	7668	961100	10939	0.982231	0.897952	0.860493
JJR	2771	256	1043957	196	0.999568	0.915428	0.933940
JJS	2937	164	1043982	97	0.999751	0.947114	0.968029
LRB	0	0	1047166	14	0.999987	NaN	0.000000
MD	6973	36	1040171	0	0.999966	0.994864	1.000000
NN	138041	6907	894474	7758	0.985996	0.952348	0.946790
NNP	127754	11728	904027	3671	0.985295	0.915917	0.972068
NNPS	1967	1234	1043425	554	0.998293	0.614495	0.780246
NNS	72706	1364	969980	3130	0.995708	0.981585	0.958727
PDT	134	55	1046978	13	0.999935	0.708995	0.911565
POS	11249	9	1035917	5	0.999987	0.999201	0.999556
PRP	13255	72	1033790	63	0.999871	0.994597	0.995270
PRP$	8614	71	1038454	41	0.999893	0.991825	0.995263
RB	19209	1361	1025567	1043	0.997704	0.933836	0.948499
RBR	833	359	1045766	222	0.999445	0.698826	0.789573
RBS	248	27	1046857	48	0.999928	0.901818	0.837838
RP	2193	315	1044375	297	0.999416	0.874402	0.880723
RRB	0	0	1047166	14	0.999987	NaN	0.000000
TO	23061	188	1023931	0	0.999820	0.991914	1.000000
UH	1	2	1047154	23	0.999976	0.333333	0.041667
VB	23348	1091	1021878	863	0.998134	0.955358	0.964355
VBD	37954	2345	1005456	1425	0.996400	0.941810	0.963813
VBG	18519	1056	1026999	606	0.998413	0.946054	0.968314
VBN	29637	1824	1013028	2691	0.995688	0.942023	0.916759
VBP	15658	739	1030283	500	0.998817	0.954931	0.969056
VBZ	24194	459	1021762	765	0.998831	0.981382	0.969350
WDT	3562	207	1043275	136	0.999672	0.945078	0.963223
WP	2523	14	1044624	19	0.999968	0.994482	0.992526
WP$	99	0	1047081	0	1.000000	1.000000	1.000000
WRB	2178	1	1044995	6	0.999993	0.999541	0.997253
``	0	0	1043490	3690	0.996476	NaN	0.000000

TagWise Evaluation Metrics for NER using NLTK

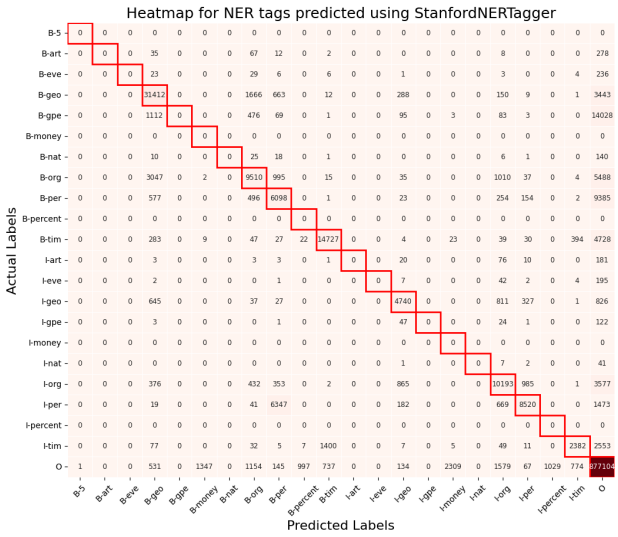
TagWise Evaluation Metrics for NER using NLTK							
	TP	FP	TN	FN	Accuracy	Precision	Recall
B-art	0	0	1048173	402	0.999617	NaN	0.000000
B-eve	0	0	1048267	308	0.999706	NaN	0.000000
B-geo	764	1126	1009805	36880	0.963755	0.404233	0.020295
B-gpe	13279	39770	992935	2591	0.959601	0.250316	0.836736
B-nat	0	0	1048374	201	0.999808	NaN	0.000000
B-org	9325	8381	1020051	10818	0.981690	0.526658	0.462940
B-per	10250	14299	1017286	6740	0.979936	0.417532	0.603296
B-tim	0	0	1028242	20333	0.980609	NaN	0.000000
I-art	0	0	1048278	297	0.999717	NaN	0.000000
I-eve	0	0	1048322	253	0.999759	NaN	0.000000
I-geo	483	786	1040375	6931	0.992640	0.380615	0.065147
I-gpe	63	3649	1044728	135	0.996391	0.016972	0.318182
I-nat	0	0	1048524	51	0.999951	NaN	0.000000
I-org	5675	2568	1029223	11109	0.986957	0.688463	0.338120
I-per	9281	3989	1027335	7970	0.988595	0.699397	0.537998
I-tim	0	0	1042047	6528	0.993774	NaN	0.000000
O	880290	44597	116070	7618	0.950204	0.951781	0.991420

StanfordCoreNLP Evaluation Plots

Heatmap for POS tags predicted using StanfordPOSTagger



Heatmap for NER tags predicted using StanfordNERTagger



TagWise Evaluation Metrics for POS using StanfordPOSTagger

TagWise Evaluation Metrics for POS using StanfordPOSTagger

	TP	FP	TN	FN	Accuracy	Precision	Recall
$	1138	2	1043628	9	0.999989	0.998246	0.992153
,	32756	495	1011525	1	0.999525	0.985113	0.999969
.	47823	8	996939	7	0.999986	0.999833	0.999854
:	221	441	1043794	321	0.999271	0.333837	0.407749
;	0	0	1044568	209	0.999800	NaN	0.000000
CC	23650	193	1020878	56	0.999762	0.991905	0.997638
CD	24185	119	1019963	510	0.999398	0.995104	0.979348
DT	97919	70	946253	535	0.999421	0.999286	0.994566
EX	645	3	1044111	18	0.999980	0.995370	0.972851
FW	0	181	1044595	1	0.999826	0.000000	0.000000
IN	117883	8933	915032	2929	0.988646	0.929559	0.975756
JJ	71747	4847	961519	6664	0.988982	0.936718	0.915012
JJR	2843	211	1041599	124	0.999679	0.930910	0.958207
JJS	2130	77	1041666	904	0.999061	0.965111	0.702044
LRB	0	0	1044776	1	0.999999	NaN	0.000000
MD	6959	7	1037797	14	0.999980	0.998995	0.997992
NN	140659	5630	893352	5136	0.989695	0.961515	0.964772
NNP	126830	6159	907221	4567	0.989734	0.953688	0.965243
NNPS	786	2530	1039726	1735	0.995918	0.237033	0.311781
NNS	73004	683	968254	2836	0.996632	0.990731	0.962605
PDT	145	66	1044564	2	0.999935	0.687204	0.986395
POS	11235	2	1033521	19	0.999980	0.999822	0.998312
PRP	13311	43	1031416	7	0.999952	0.996780	0.999474
PRP$	8631	4	1036118	24	0.999973	0.999537	0.997227
RB	19268	2932	1021593	984	0.996252	0.867928	0.951412
RBR	831	321	1043401	224	0.999478	0.721354	0.787678
RBS	261	828	1043653	35	0.999174	0.239669	0.881757
RP	2296	455	1041832	194	0.999379	0.834606	0.922088
RRB	0	0	1044776	1	0.999999	NaN	0.000000
TO	14710	10	1021706	8351	0.991997	0.999321	0.637873
UH	23	39	1044714	1	0.999962	0.370968	0.958333
VB	23586	601	1019965	625	0.998827	0.975152	0.974185
VBD	38436	1319	1004079	943	0.997835	0.966822	0.976053
VBG	18198	763	1024889	927	0.998382	0.959760	0.951529
VBN	30666	2322	1010127	1662	0.996187	0.929611	0.948589
VBP	15846	344	1028275	312	0.999372	0.978752	0.980691
VBZ	24702	294	1019523	258	0.999472	0.988238	0.989663
WDT	3649	287	1040792	49	0.999678	0.927083	0.986750
WP	2514	8	1042227	28	0.999966	0.996828	0.988985
WP$	99	0	1044678	0	1.000000	1.000000	1.000000
WRB	2183	0	1042593	1	0.999999	1.000000	0.999542

TagWise Evaluation Metrics for NER using StanfordNERTagger

TagWise Evaluation Metrics for NER using StanfordNERTagger

	TP	FP	TN	FN	Accuracy	Precision	Recall
B-art	0	0	1042419	402	0.999615	NaN	0.000000
B-eve	0	0	1042513	308	0.999705	NaN	0.000000
B-geo	31412	6743	998434	6232	0.987558	0.823273	0.834449
B-gpe	0	0	1026954	15867	0.984785	NaN	0.000000
B-nat	0	0	1042620	201	0.999807	NaN	0.000000
B-org	9510	4505	1018175	10631	0.985486	0.678559	0.472171
B-per	6098	8672	1017159	10892	0.981239	0.412864	0.358917
B-tim	14727	2178	1020364	5552	0.992587	0.871162	0.726219
I-art	0	0	1042524	297	0.999715	NaN	0.000000
I-eve	0	0	1042568	253	0.999757	NaN	0.000000
I-geo	4740	1709	1033698	2674	0.995797	0.734998	0.639331
I-gpe	0	0	1042623	198	0.999810	NaN	0.000000
I-nat	0	0	1042770	51	0.999951	NaN	0.000000
I-org	10193	4810	1021227	6591	0.989067	0.679397	0.607305
I-per	8520	1639	1023931	8731	0.990056	0.838665	0.493884
I-tim	2382	1185	1035120	4134	0.994899	0.667788	0.365562
O	877104	46694	113902	5121	0.950313	0.949454	0.994195

NER Performance Analysis

- **Geographical Entities (B-geo)**: Stanford NER shines with a precision of 82.33% and a recall of 83.44%, indicating a strong ability to recognize geographical names accurately. This outperforms NLTK's precision of 40.42% and recall of 2.03%, and spaCy's precision of 56.50% and recall of 4.66%, showing Stanford NER's superior capability in this category.

  Across the tools, B-geo misclassifications into B-gpe were notable, with NLTK misclassifying 29,235 instances, spaCy with 28,376, and Stanford less prone to this specific error but more likely to classify B-geo as O (3,443 instances). This suggests that while NLTK and spaCy struggle to differentiate between geographical and geopolitical entities, Stanford's model is more conservative, potentially missing geographical entities altogether rather than confusing them with geopolitical ones.

- **Geopolitical Entities (B-gpe)**: NLTK demonstrates a significant recall of 83.67% but with a lower precision of 25.03%, suggesting it is really good at identifying geopolitical entities, but with more false positives. Stanford NER did not provide precision for B-gpe, showing a recall of 0.00%, and spaCy's metrics show a precision of 3.04% and recall of 7.07%, indicating a conservative but less effective approach.
- **Organizations (B-org)**: Stanford NER exhibits the highest precision among the three at 67.86%, indicating its strong ability to accurately identify organizations. Its recall is 47.22%, compared to NLTK's precision of 52.67% and recall of 46.29%, and spaCy's precision of 24.90% and recall of 42.95%, showing Stanford's advantage in recognizing organizations accurately.

  Organizational entities posed challenges for all models, particularly with NLTK and Stanford misclassifying B-org as O in a significant number of instances (1,843 for NLTK and a similar pattern for Stanford). spaCy, while also showing misclassifications into O, had a unique issue of confusing B-org with B-gpe, reflecting its attempt to leverage context but sometimes overgeneralizing from it.

- **Persons (B-per)**: Stanford NER and NLTK show competitive performance in identifying person names, with Stanford having a precision of 41.29% and a recall of 35.89%. NLTK's precision is 41.75% with a higher recall of 60.33%. SpaCy offers a precision of 38.98% and recall of 39.40%, indicating a balanced approach but with room for improvement.

Person name recognition showed varied performance, with NLTK incorrectly tagging B-per as B-org in several instances, indicating confusion between personal and organizational names. spaCy and Stanford showed better performance in correctly identifying B-per tags, though they also had errors, occasionally misclassifying person names as O due to possibly ambiguous contexts.

- **Time Expressions (B-tim and I-tim)**: Stanford NER demonstrates exceptional precision at 87.12% and recall of 72.62% for B-tim, indicating its strong capability in recognizing temporal expressions. SpaCy also performs well with a precision of 61.57% and a high recall of 75.13% for B-tim, showcasing its effectiveness in temporal entity recognition.
- **Nested or Inside Entities (I-geo, I-org, I-per)**: Stanford NER generally shows higher precision across these categories, indicating its strength in accurately identifying entities within larger contexts. For instance, it offers a precision of up to 83.87% (I-per) and a recall of up to 60.73% (I-org). SpaCy's precision peaks at 79.87% (I-per) with a recall up to 63.51% (I-org), indicating its capability in recognizing nested entities.
- **B-art, B-eve, B-nat, I-art, I-eve, I-nat**: These entities are challenging for all libraries with generally low precision and recall. However, spaCy shows relatively higher recall for rare events (B-eve and I-eve), such as 59.68% for I-eve, suggesting it is better at recognizing these when they occur.
- **Non-entity Segments (O)**: Stanford NER and NLTK demonstrate high precision (94.95% and 95.18%, respectively) and recall (99.42% and 99.14%), indicating their strength in correctly identifying segments that do not represent named entities. SpaCy also performs well in this category, with a precision of 97.91% and recall of 95.84%.

Overall, Stanford NER tends to offer high precision across most entity types, especially for more specific entity recognition tasks like time expressions and nested entities. NLTK, while demonstrating high recall in certain categories, often has lower precision, suggesting a tendency to identify more entities but with a higher rate of false positives. SpaCy provides a balanced performance with notable strengths in recognizing temporal entities and nested organizations but shows room for improvement in precision for some entity types.

POS Performance Analysis

Symbols and Punctuation

- **Dollar Sign ($), Comma (,), Period (.), and Colon (:)**: Libraries demonstrate high precision and recall for the dollar sign, comma, and period, with NLTK often achieving perfect metrics (100% precision and recall for periods). For colons, Stanford's precision dips to 33.38% with a recall of 40.77%, highlighting its comparative struggle with this punctuation. In contrast, spaCy improves significantly with a precision of 44.30% and recall of 76.32% for colons, suggesting its better handling of complex punctuation.

Parts of Speech

- **Conjunctions (CC), Cardinal Numbers (CD), Determiners (DT), Existentials (EX)**: Stanford excels, showing nearly perfect precision for determiners at 99.93% and recall for existentials at 97.29%. NLTK and spaCy also perform well but with slight differences; spaCy's precision for cardinal numbers is 99.51%, slightly higher than NLTK's, indicating nuanced strengths in numerical identification.
- **Foreign Words (FW), Prepositions (IN), Adjectives (JJ, JJR, JJS)**: Stanford stands out in adjective identification with a precision of 93.67% for JJ tags, compared to spaCy's higher precision of 93.46% for JJR, showcasing each library's capability in distinguishing subtle differences in adjective forms.
- **Modal Verbs (MD), Nouns (NN, NNP, NNPS, NNS)**: Across libraries, modal verbs show high precision and recall, with Stanford achieving a 99.90% precision for MD. For proper nouns (NNP), Stanford's recall is 96.52%, indicating its strength in recognizing named entities, while spaCy's precision for singular nouns (NN) at 96.61% highlights its accuracy in noun identification.
- **Predeterminers (PDT), Possessive Endings (POS), Personal Pronouns (PRP, PRP$)**: Exceptional metrics are observed, with Stanford and spaCy showing a precision of 99.95% and 99.93%, respectively, for PRP$, emphasizing their accuracy in identifying possessive pronouns.
- **Adverbs (RB, RBR, RBS), Particles (RP)**: SpaCy's nuanced understanding of adverbial forms is evident with a precision of 86.45% for RB, surpassing Stanford's and NLTK's performance. However, Stanford maintains a high recall for particles (RP) at 92.21%, shows its comprehensive identification capabilities.
- **Infinitival "to" (TO), Interjections (UH)**: While Stanford and NLTK show high precision for TO, spaCy's performance in recognizing interjections with a

precision of 21.82% and recall of 100% indicates its sensitivity to nuanced language elements.

- **Verbs (VB, VBD, VBG, VBN, VBP, VBZ)**: All libraries excel in verb form identification, with Stanford generally achieving high recall. SpaCy, however, distinguishes itself with a higher precision in past participle (VBN) forms at 97.16%, underscoring its precision in verb form distinction.
- **Wh-elements (WDT, WP, WP$, WRB)**: SpaCy's handling of Wh-elements is particularly effective, with a precision of 98.50% for WP, indicating its really good in interrogative and relative structure identification.

Quotation Marks

- **Opening Quotation Marks (``)**: SpaCy's exceptional precision in this category, where Stanford and NLTK did not identify any instances, showcases its capability in recognizing direct speech and quotations within text.

Overall, while Stanford, NLTK, and spaCy each exhibit strong performance across a broad spectrum of POS tags, their specific strengths in precision and recall for different categories highlight their suitability for varied POS tagging needs.

Q2:
I have used Doccano for the manual tagging of POS and NER tags, taking each word as a token. After making amendments to the codes, we observed a significant drop in tag-wise accuracy, precision, and recall for both POS and NER tagging for all the 3 libraries.

a. For example, whilst using the NLTK library and performing a NER with it, for the O tag which was the most commonly used tag, the accuracy dropped to 0.78, the precision was 0.72 and the recall was 0.75, which was a significant drop from the original dataset where we had an accuracy of 0.95, a precision of 0.95 and a recall of 0.99 thus highlighting the issue. Additionally, for less common tags like B-org, the accuracy came down even more from 0.98 to 0.81 whilst precision was down to 0.21 and recall was at 0.15. Whilst using POS tagging in terms of precision, NN in the original dataset had a precision of 0.95 whilst in the manually annotated POS tagging, the precision was 0.72, whilst recall was 0.75 and accuracy was 0.82. For less common tags like PRP, we had achieved an accuracy of 0.99, a recall of 0.98 and precision of 0.99 whilst here it came down to 0.84 (accuracy), 0.88 (precision), 0.83 (recall) hence showcasing the drop

b. For the stanford core nlp package likewise we saw a downgrade too: For the original dataset in Q1, we achieved great result in terms of accuracy, precision and recall across NER and POS tagging however manual annotation had stark differences especially for example between tags like NN and NNS or VB and VBZ or VB and VBG

Tags	Accuracy (original)	Accuracy (new)	Precision (original)	Precision (new)	Recall (original)	Recall (new)
O	0.95	0.78	0.94	0.72	0.99	0.75
B-org	0.98	0.81	0.68	0.21	0.47	0.15
NN	0.96	0.82	0.97	0.79	0.96	0.75
PRP	0.99	0.84	0.99	0.88	0.99	0.83

c. Whilst using spacy also manual annotation revealed significantly poor results too

Tags	Accuracy (original)	Accuracy (new)	Precision (original)	Precision (new)	Recall (original)	Recall (new)
O	0.95	0.78	0.98	0.72	0.96	0.75
B-org	0.98	0.81	0.42	0.21	0.24	0.15
NN	0.98	0.82	0.96	0.79	0.95	0.75
PRP	0.99	0.84	0.99	0.88	0.99	0.83

In conclusion, we observed that manual annotation had several pitfalls for example difference in labelling since we were not experts which could be a part explainer of the results difference