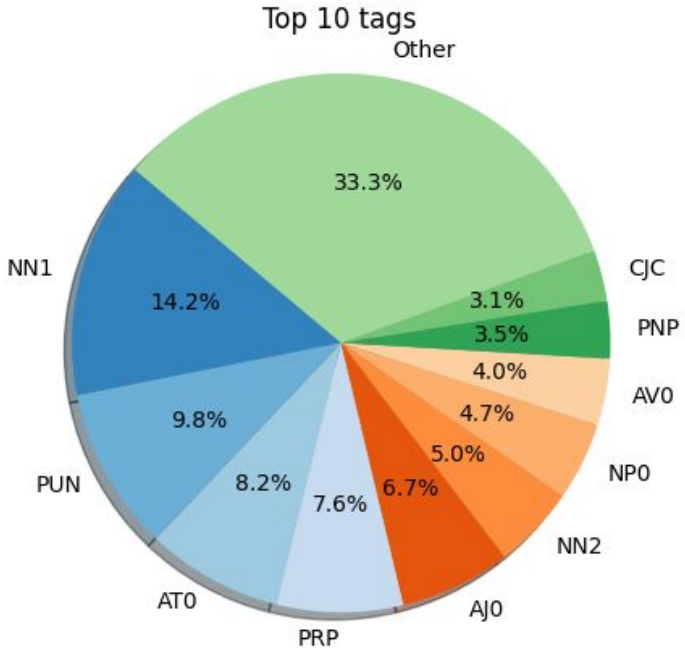
AI Project - POS Tagging

Team Members:

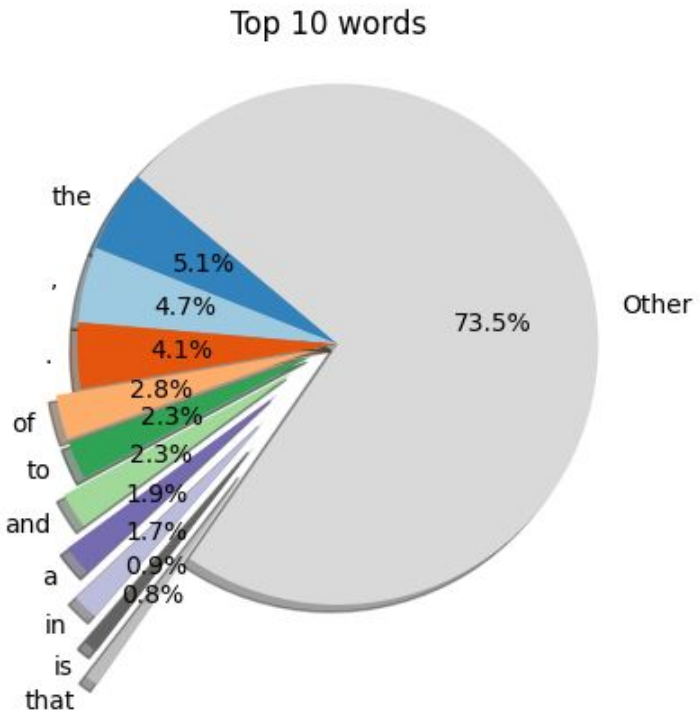
- Katkar Prathamesh Shivaji (18114038)
- Ritesh Singh (18114067)
- Utkarsh (18114080)

Top 10 frequent tags



% Part	Word
14.2%	NN1
9.8%	PUN
8.2%	AT0
7.6%	PRP
6.7%	AJ0
5.0%	NN2
4.7%	NP0
4.0%	AV0
3.5%	PNP
3.1%	CJC

Top 10 frequent words



% Part	Word
5.1%	the
4.7%	,
4.1%	.
2.8%	of
2.3%	to
2.3%	and
1.9%	a
1.7%	in
0.9%	is
0.8%	that

Analysis

1. Every sentence contains a subject and a predicate. In the dataset, NN1 is used more frequently as a subject than others so it is the most frequent tag.
2. Every sentence includes some punctuation marks so PUN should be among the top tags.
3. Period(.) is behind comma(,) in word frequency which indicates that most of the sentences are compound sentences or contain comma(,) separated lists in the dataset. These sentences contains more comma(,) than period(.)
4. AT0's 3rd place is justified because most of them occurs together with NN1 (example: a train, the book, an apple, etc) and NN1 is the most frequently used tag.
5. A simple predicate contains the verb and can also contain modifying words, phrases, or clauses but still no verb is present in top 10 frequent tags because in the c5 POS tagset verbs are further distributed into 25 tags almost evenly.