# Exercise 1

## Part a

*How is the result compared to using the full brown tagset in the introduction?  
Why do you think one of the tagsets yields higher scores than the other one?*

Result from the run, using the Universal Tagset instead of the Brown Tagset: **0.8689**

Result from the previous run with the given Brown Tagset: 0.7915

The large difference in the scores achieved by using two different tagsets might be a result of the different tags applied and their number, too. The Brown tagset uses 87 different tags to classify the parts of speech it encounters and therefore it is a very elaborate tagset. On the other hand, the Universal Tagset only applies 12 tags, which makes it a more reduced tool to use. However, this reduction to 12 “core” tags for classifying parts of speech might lead to higher accuracies in classification tasks, since the tagger is then less prone to wrong tagging of words, or – to put it differently – the probability that the tagger tags a word correctly is higher when there are only 12 tags then when there are 87 tags to choose from.

## Part b

Result of the baseline tagger: 0.9296

The baseline tagger as it is implemented in Part b of this exercise already achieves a very high accuracy of 93% and therefore outperforms the tagger from Part a by approximately 6%.

However, this comes as surprise (to me), since the baseline tagger uses a very rudimental method to decide on the tags. It decides which POS-tag to choose for a particular word on the basis of the most frequently seen tag for that word in the training set, which does not seem to be a very elaborate method. Yet, the tagger from Part a also “only” looks at the suffixes of a word and whether it is the first word in a sentence or not, which does not seem to be good indicators for deciding on the POS-tags, either.

# Exercise 2

## Part a

*Train the ScikitConsecutivePosTagger on the \*news\_train\* set and test on the \*news\_dev\_test\* set with the \*pos\_features\*. Do you get the same result as with the same data and features and the NLTK code in exercise 1a?*

Results from running the code on the ScikitConsecutivePosTagger: 0.857

The result when running the same data

features with the ScikitConsecutivePosTagger differ slightly from the ones achieved with the NLTK code. To be more precise, the tagger implemented in exercise 2a achieves a slightly lower accuracy than the one implemented in exercise 1a.

## Part b

With the best choice of alpha, do you get the same results as with the NLTK code in exercise 1a, worse results or better results?

|  |  |
| --- | --- |
| **Value of alpha-parameter with BernoulliNB** | **Result** |
| 1 | 0.857 |
| 0.5 | 0.8749 |
| 0.1 | 0.8695 |
| 0.01 | 0.8683 |
| 0.001 | 0.8651 |
| 0.0001 | 0.8631 |

The best choice of the alpha-parameter is alpha=0.5, which yields an accuracy of 0.8749. This is slightly better than what the NLTK-code gives as a result. Moreover, the parameter alpha=0.1 yields a better result as well, while all the other parameters fall behind the NLTK-result from exercise 1a.

## Part c

Did the extended feature selector beat the baseline (0.9296)?

The extended version of the feature selector only beat the baseline when using alpha-values of 0.01 and smaller. Yet, the baseline only performs slightly worse and is not beaten extremely by the extended feature selector and tagger.

The extended feature selector should intuitively beat the baseline because for the baseline calculation a very rudimental method was used, by assigning the words’ most common POS-tag to the respective word. Therefore, the baseline tagger doesn’t take a word’s context into account. However, the context (or other features, for that matter) can be a very helpful indicator when it comes to POS-tagging as it might help to decide which POS is the correct one.   
By looking at the results of this comparison, however, one can conclude that the rudimentary baseline tagger is already a good approach to assigning POS-tags to words, which achieves good results when tagging the texts.   
That both taggers achieve almost equally highly could be due to the fact that – after extending the NLTK feature selector - both of them look at the word to be classified now. The extended feature selector only additionally takes the previous word and the suffixes of the word into account. Those features do not seem to be really strong indicators of the POS-tag a particular word should receive (according to the results of this run, at least (I would intuitively think that the previous word has a stronger impact)).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Value of alpha-parameter with BernoulliNB** | **Feature extractor from Part b: Results** | **Results from extended feature extractor** | **Improvement of accuracy from Part b to Part c?** | **Baseline beaten?** |
| 1 | 0.857 | 0.8874 |  |  |
| 0.5 | 0.8749 | 0.9166 |  |  |
| 0.1 | 0.8695 | 0.9244 |  |  |
| 0.01 | 0.8683 | 0.9303 |  |  |
| 0.001 | 0.8651 | 0.933 |  |  |
| 0.0001 | 0.8631 | 0.934 |  |  |

Running this experiment with the extended feature extractor and various alpha-values, the best results are achieved when using alpha=0.0001. In general, the extended feature extractor performs better than the not-extended one which only looks at the previous word instead of the previous and current word.

# Exercise 3

## Part a

Result of running with the extended feature extractor from the previous Exercise 2 in combination with Logistic Regression as classification algorithm: 0.9514

This result is even better than the best result with the Naïve Bayes classifier (best one here: 0.934).

## Part b

|  |  |
| --- | --- |
| **Value of C-parameter for smoothing in Logistic Regression (regularization)** | **Result** |
| 0.01 | 0.8499 |
| 0.1 | 0.9265 |
| 1.0 | 0.9514 |
| 10.0 | 0.9545 |
| 100.0 | 0.9531 |
| 1000.0 | 0.954 |

The classification with the logistic regression and various parameters for regularization yields the best results when using C=10.0. As smaller C values stand for stronger regularization, the best results are achieved with medium regularization in this case here. With the smallest C-value, the worst result is obtained with only 85% of accuracy in the POS-tags.

# Exercise 4

## Part a

Result with added feature in feature extractor – the next word in the sentence – and the best C from Exercise 3: 0.9651, which is a further improvement in accuracy

## Part b

|  |  |
| --- | --- |
| **Features observed by feature extractor** | **Accuracy** |
| Including all previous features plus checking whether word is a number | 0.9658 |
| Including all previous features plus checking whether word is capitalized | 0.9646 |
| Combining the previous two settings | 0.9646 |
| Checking whether all letters in word are capitalized, or only the first or none | 0.9644 |
| Previous setting + checking if string is a number | 0.9644 |

The best results can be obtained by checking whether the string is a number and also considering the following, the previous and the word itself.

Speed of running the experiment: 25sek

# Exercise 5

## Part b

Baseline Tagger Accuracy: 0.9478

## Part c

Accuracy of the tagger with the whole data set and the best settings from the previous exercises: 0.969  
Speed of running the experiment: 3:16 min

# Exercise 6

## Part a

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **.** | **ADJ** | **ADP** | **ADV** | **CONJ** | **DET** | **NOUN** | **NUM** | **PRON** | **PRT** | **VERB** | **X** |
| **.** | 12724 | . | . | . | . | . | . | . | . | . | . | . |
| **ADJ** | . | 6539 | 2 | 161 | . | . | 374 | 2 | . | 4 | 162 | . |
| **ADP** | 1 | 11 | 12547 | 70 | 5 | 16 | 8 | . | 15 | 189 | 17 | . |
| **ADV** | . | 211 | 113 | 4406 | 9 | 8 | 47 | . | . | 20 | 17 | . |
| **CONJ** | . | . | . | 7 | 3225 | 1 | . | . | . | . | 1 | . |
| **DET** | . | 1 | 34 | 14 | 3 | 12083 | 13 | . | 26 | . | . | . |
| **NOUN** | . | 294 | 5 | 39 | . | 3 | 23724 | 41 | 2 | 7 | 244 | 7 |
| **NUM** | . | 5 | 1 | . | . | . | 36 | 1242 | . | . | 2 | . |
| **PRON** | . | . | 49 | 3 | . | 33 | 9 | . | 3983 | . | 1 | 1 |
| **PRT** | . | 3 | 102 | 10 | . | 2 | 35 | . | 2 | 2387 | 3 | . |
| **VERB** | . | 83 | 11 | 16 | . | 2 | 426 | . | . | 4 | 15317 | . |
| **X** | 5 | 5 | 4 | 1 | . | 3 | 74 | . | 3 | . | 4 | 28 |

(row = reference; col = test)

## Part b

|  |  |  |  |
| --- | --- | --- | --- |
| **Tag** | **Precision** | **Recall** | **F-measure** |
| **.** | 0.9995 | 1.00 | 0.9998 |
| **ADJ** | 0.9143 | 0.9027 | 0.9084 |
| **ADP** | 0.9751 | 0. 9742 | 0. 9746 |
| **ADV** | 0.9321 | 0.9120 | 0.9220 |
| **CONJ** | 0.9948 | 0.9972 | 0.9960 |
| **DET** | 0.9944 | 0.9925 | 0.9935 |
| **NOUN** | 0.9587 | 0.9737 | 0.9661 |
| **NUM** | 0.9665 | 0.9658 | 0.9662 |
| **PRON** | 0.9881 | 0.9765 | 0.9822 |
| **PRT** | 0.9142 | 0.9383 | 0.9261 |
| **VERB** | 0.9714 | 0.9658 | 0.9686 |
| **X** | 0.7778 | 0.2205 | 0.3436 |

## Part c

|  |  |
| --- | --- |
| Macro-Precision | 0.9489083333333331 |
| Macro-Recall | 0.9015999999999998 |
| Macro-F-Measure | 0.9122583333333334 |

# Exercise 7

|  |  |  |
| --- | --- | --- |
| TOKEN PRED GOLD  ---------- ------ ------  He PRON PRON  assured VERB VERB  Mr. NOUN NOUN  Martinelli NOUN NOUN  and CONJ CONJ  the DET DET  council NOUN NOUN  that DET ADP  he PRON PRON  would VERB VERB  study NOUN VERB  the DET DET  correct ADJ ADJ  method NOUN NOUN  and CONJ CONJ  report NOUN VERB  back ADV ADV  to PRT ADP  the DET DET  council NOUN NOUN  as ADP ADV  soon ADV ADV  as ADP ADP  possible ADJ ADJ  . . . | TOKEN PRED GOLD  ---------- ------ ------  Mr. NOUN NOUN  Martinelli NOUN NOUN  said VERB VERB  yesterday NOUN NOUN  that DET ADP  the DET DET  Citizens NOUN NOUN  Group NOUN NOUN  of ADP ADP  Johnston NOUN NOUN  will VERB VERB  meet NOUN VERB  again ADV ADV  July NOUN NOUN  24 NUM NUM  to PRT PRT  plan NOUN VERB  further ADV ADJ  strategy ADJ NOUN  in ADP ADP  the DET DET  charter NOUN NOUN  movement NOUN NOUN  . . . | TOKEN PRED GOLD  ----------- ------ ------  He PRON PRON  said VERB VERB  that DET ADP  the DET DET  group NOUN NOUN  has VERB VERB  no DET DET  candidates NOUN NOUN  for ADP ADP  the DET DET  charter NOUN NOUN  commission NOUN NOUN  in ADP ADP  mind NOUN NOUN  at ADP ADP  present ADV ADJ  , . .  but CONJ CONJ  that DET ADP  it PRON PRON  will VERB VERB  undoubtedly ADV ADV  endorse ADJ VERB  candidates NOUN NOUN  when ADV ADV  the DET DET  time NOUN NOUN  comes VERB VERB  . . . |
| TOKEN PRED GOLD  --------- ------ ------  `` . .  After ADP ADP  inspiring VERB VERB  this DET DET  , . .  I PRON PRON  think VERB VERB  we PRON PRON  should VERB VERB  certainly ADV ADV  follow VERB VERB  through ADP PRT  on ADP ADP  it PRON PRON  '' . .  , . .  he PRON PRON  declared VERB VERB  . . . | TOKEN PRED GOLD  -------------- ------ ------  `` . .  It PRON PRON  has VERB VERB  become VERB VERB  our DET DET  responsibility NOUN NOUN  and CONJ CONJ  I PRON PRON  hope NOUN VERB  that DET ADP  the DET DET  Citizens NOUN NOUN  Group NOUN NOUN  will VERB VERB  spearhead ADJ VERB  the DET DET  movement NOUN NOUN  '' . .  . . . |  |

*Identify the words that are tagged differently. Comment on each of the differences.  
Would you say that the predicted tag is wrong? Or is there a genuine ambiguity such that both answers are defendable?  
Or is even the gold tag wrong?*

What is perceptible from the five examples above is that sometimes, verbs are mistaken for nouns when their form is equal to a noun as in (to) plan – (the) plan / (to) hope – (the) hope / (to) report – (the) report. This is a special case where the context, or more specifically the previous word might help to determine which POS is correct. If there is a “to” before the word, one can conclude that it is a verb. However, “to” does not always precede a verb, therefore it might not be the best indicator either. In the examples above, the verbs mistaken for a noun are not preceded by “to”.

Furthermore, it is striking that the word “that” is wrongly tagged in all occurrences in the five examples above. It is predicted to be a determiner (DET), yet, the gold tag defines it as adposition (ADP). In the [Universal Tagset definition](https://universaldependencies.org/u/pos/ADP.html) of POS-tags it says

*“Adposition is a cover term for prepositions and postpositions. Adpositions belong to a closed set of items that occur before (preposition) or after (postposition) a complement composed of a noun phrase, noun, pronoun, or clause that functions as a noun phrase, and that form a single structure with the complement to express its grammatical and semantic relation to another unit within a clause.”*

Therefore, the gold tag might come from the indication that “that” is usually used in front of clauses or (noun) phrases. The predicted tag DET-determiner for “that” might come from the use of the word as

*“a determiner, a demonstrative pronoun and a relative pronoun”,*

as the [Cambridge Dictionary](https://dictionary.cambridge.org/grammar/british-grammar/that) defines it. Looking at the examples, I would say that the gold tag is the correct one, as one cannot replace “that” with “this”, for example, which would argue in favor of “that” acting as determiner. It seems that the tagger is not trained well enough to solve this ambiguity of POS-tags.

Concerning other mistakes: I cannot explain why the tagger tagged “strategy” as adjective-ADJ and not as noun. The proper adjective to the noun would be “strategic” or “strategical”, which are both different from the word which is given in the example here. Moreover, the tagger assigned the word “spearhead” to be an adjective-ADJ instead of a verb. What is confusing here is that the word is not wrongly tagged as noun, as “spearhead” is a proper noun in English as well. Making this mistake would have suited the Noun-Adjective-ambivalence which was described earlier in this exercise. Yet, assigning the word the the Adjective-tag does not really make sense.

As last mistakes, it is perceptible that in the examples above, ADP was mistaken for PRT once and vise versa as well. Looking on the examples the [nltk-book (Ch. 5.2.3)](https://www.nltk.org/book/ch05.html) gives for both categories, we can see that there are some words falling into both. Given that, it is understandable why the tagger might make mistakes in assigning the categories, as they do not seem to be clearly separated from each other.

# Exercise 8

## Part a

Result of running the tagger on the *test* data set: 0.9698

The result of running the tagger on the *dev\_test* data set was 0.969 (see Ex5, Part c), therefore there is hardly any difference between the accuracies achieved on testing on both data sets. Testing on the *test* data set achieves a slightly better result than the other experiment, however, this difference should not play a major role in assessing the quality of the tagger.

## Part b

Result of testing the tagger on the *adventure* set: 0.963

Result of testing the tagger on the *hobbies* set: 0.9546

As we can see, the tagger does not perform as highly on these two data sets as on the *test* data set in Part a before. A reason for that could be that before, we trained the tagger on the same genres of text we would later test it on, therefore the contents of the test set do not hugely differ from the ones on which the tagger was trained in the exercises before. The two genres *adventure* and *hobbies* were left out when creating the training test sets, therefore these two genres are completely new to the tagger and therefore it might not perform as good on them.

It is also perceptible from the results of running the tagger on the *adventure* and on the *hobbies* set, that the tagger performs better on the further than on the latter. A reason for that could be that the genre *adventures* consist of novels and short stories (see <http://icame.uib.no/brown/bcm.html>), which are both included in other genres as well – e.g. *humor, romance, science fiction*. Furthermore, the Brown Corpus includes 126 examples of *imaginative* *prose* to which *adventure* can be counted as well. This leads to a well-trained tagger when it comes to this whole genre. Therefore, the tagger might have been more familiar with this type of writing and therefore was able to better tag the words with a respective POS.

However, the genre *hobbies* consists of material from books and periodicals and belongs to the sub-category of informative prose. This sub-category consists only of 17 examples, which is really few compared to the sub-category of *imaginative prose*. Of course, book texts might be longer, however, the training material for this sub-category is still less than for the other sub-category.

# Exercise 9

## Part a

Result of training the HMM tagger on *news­\_train­* and running it on *news\_test*: 0.8995

Result of training on *train* and running HMM tagger on *test*: 0.9518  
Speed of this experiment: 2:50 mins

Compared to our first tagger, the HMM tagger performs a bit worse with an accuracy of only 95% compared to the 97% achieved before. In terms of speed, the HMM tagger works slightly faster than the tagger which was used for the previous experiments. It runs about half a minute less in time, however, it also performs slightly worse than the previous tagger so this in not really an advantage here.

## Part b

Result of training the perceptron tagger on *news­\_train­* and running it on *news\_test*: 0.9638

Result of training on *train* and running the perceptron tagger on *test*: 0.9793  
Speed of the experiment: 3:54 mins

Running the perceptron tagger achieved higher results than running the best tagger from the previous exercises. It outperforms the previous tagger by 1% with accuracy values of 0.9793 compared to 0.969 (Ex 5, Part C). With an accuracy of almost 98%, this tagger can be evaluated of being really good. As one would expect, the tagger performed better when trained on the larger *train* and *test* sets than when trained and tested on the news sets. This is due to the higher number of examples the tagger can look at in the training sets and therefore it is able to perform better.

When it comes to speed, the perceptron tagger runs about half a minute longer on the whole *train/test data set* than the tagger used in Exercise 5. Since the perceptron tagger achieves better, the 30 seconds are a reasonable “downside”.

|  |  |
| --- | --- |
| [(('Mr.', 'NOUN'),  ('Martinelli', 'NOUN'),  ('said', 'VERB'),  ('yesterday', 'NOUN'),  ('that', 'DET'),  ('the', 'DET'),  ('Citizens', 'NOUN'),  ('Group', 'NOUN'),  ('of', 'ADP'),  ('Johnston', 'NOUN'),  ('will', 'VERB'),  ('meet', 'NOUN'),  ('again', 'ADV'),  ('July', 'NOUN'),  ('24', 'NUM'),  ('to', 'PRT'),  ('plan', 'NOUN'),  ('further', 'ADV'),  ('strategy', 'ADJ'),  ('in', 'ADP'),  ('the', 'DET'),  ('charter', 'NOUN'),  ('movement', 'NOUN'),  ('.', '.')), | GOLD SENT [[('Mr.', 'NOUN'),  ('Martinelli', 'NOUN'),  ('said', 'VERB'),  ('yesterday', 'NOUN'),  ('that', 'ADP'),  ('the', 'DET'),  ('Citizens', 'NOUN'),  ('Group', 'NOUN'),  ('of', 'ADP'),  ('Johnston', 'NOUN'),  ('will', 'VERB'),  ('meet', 'VERB'),  ('again', 'ADV'),  ('July', 'NOUN'),  ('24', 'NUM'),  ('to', 'PRT'),  ('plan', 'VERB'),  ('further', 'ADJ'),  ('strategy', 'NOUN'),  ('in', 'ADP'),  ('the', 'DET'),  ('charter', 'NOUN'),  ('movement', 'NOUN'),  ('.', '.')]] |