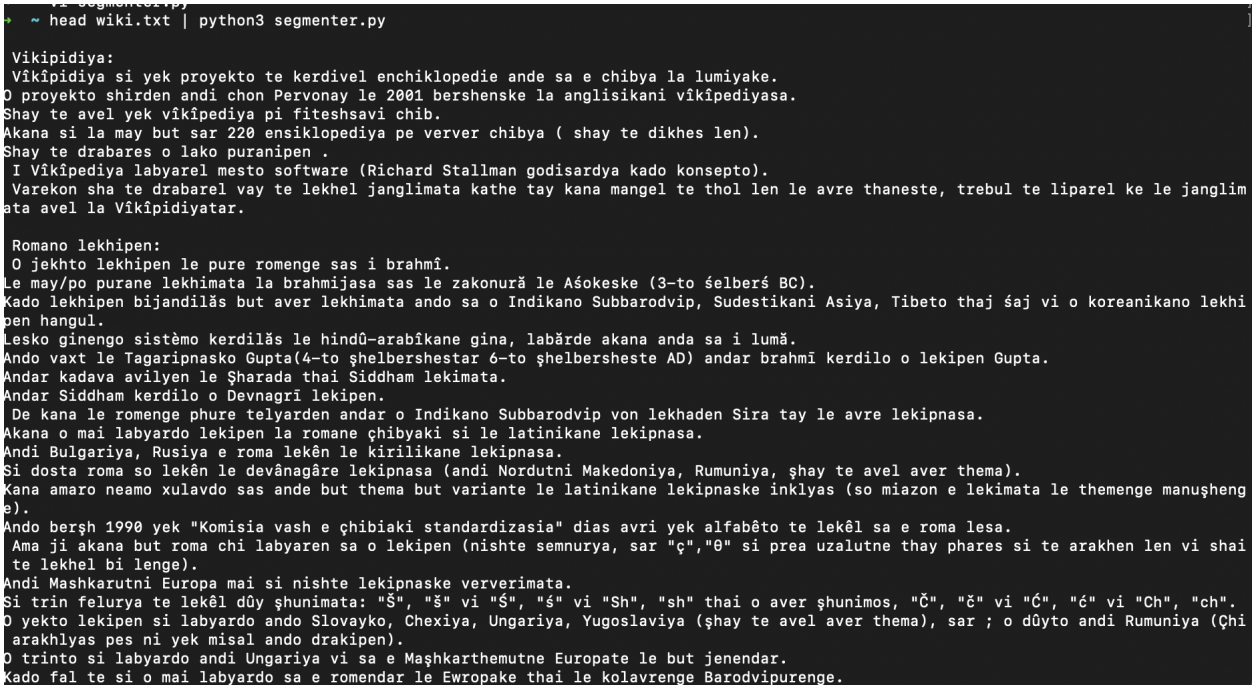
LING L-545/CSCI B-659: Computation and Linguistic analysis

Name: Manasi Swaminathan

Username: mswamina

Segmentation and Tokenization Assignment

Output for segmenter.py for Romani language



```
→  ~ head wiki.txt | python3 segmenter.py

 Vikipidiya:
 Vîkîpidiya si yek proyekto te kerdivel enchiklopedie ande sa e chibya la lumiyake.
O proyekto shirden andi chon Pervonay le 2001 bershenske la anglisikani vîkîpediyasa.
Shay te avel yek vîkîpediya pi fiteshsavi chib.
Akana si la may but sar 220 ensiklopediya pe verver chibya ( shay te dikhes len).
Shay te drabares o lako puranipen .
 I Vîkîpediya labyarel mesto software (Richard Stallman godisardya kado konsepto).
 Varekon sha te drabarel vay te lekhel janglimata kathe tay kana mangel te thol len le avre thaneste, trebul te liparel ke le janglim
ata avel la Vîkîpidiyatar.

 Romano lekhipen:
 O jekhto lekhipen le pure romenge sas i brahmî.
Le may/po purane lekhimata la brahmijasa sas le zakonură le Aśokeske (3-to śelberś BC).
Kado lekhipen bijandilăs but aver lekhimata ando sa o Indikano Subbarodvip, Sudestikani Asiya, Tibeto thaj śaj vi o koreanikano lekhi
pen hangul.
Lesko ginengo sistèmo kerdilăs le hindû-arabîkane gina, labărde akana anda sa i lumă.
Ando vaxt le Tagaripnasko Gupta(4-to şhelbershestar 6-to şhelbersheste AD) andar brahmī kerdilo o lekipen Gupta.
Andar kadava avilyen le Şharada thai Siddham lekimata.
Andar Siddham kerdilo o Devnagrī lekipen.
 De kana le romenge phure telyarden andar o Indikano Subbarodvip von lekhaden Sira tay le avre lekipnasa.
Akana o mai labyardo lekipen la romane çhibyaki si le latinikane lekipnasa.
Andi Bulgariya, Rusiya e roma lekên le kirilikane lekipnasa.
Si dosta roma so lekên le devânagâre lekipnasa (andi Nordutni Makedoniya, Rumuniya, şhay te avel aver thema).
Kana amaro neamo xulavdo sas ande but thema but variante le latinikane lekipnaske inklyas (so miazon e lekimata le themenge manuşheng
e).
Ando berşh 1990 yek "Komisia vash e çhibiaki standardizasia" dias avri yek alfabêto te lekêl sa e roma lesa.
 Ama ji akana but roma chi labyaren sa o lekipen (nishte semnurya, sar "ç","θ" si prea uzalutne thay phares si te arakhen len vi shai
 te lekhel bi lenge).
Andi Mashkarutni Europa mai si nishte lekipnaske ververimata.
Si trin felurya te lekêl dûy şhunimata: "Š", "š" vi "Ś", "ś" vi "Sh", "sh" thai o aver şhunimos, "Č", "č" vi "Ć", "ć" vi "Ch", "ch".
O yekto lekipen si labyardo ando Slovayko, Chexiya, Ungariya, Yugoslaviya (şhay te avel aver thema), sar ; o dûyto andi Rumuniya (Çhi
 arakhlyas pes ni yek misal ando drakipen).
O trinto si labyardo andi Ungariya vi sa e Maşhkarthemutne Europate le but jenendar.
Kado fal te si o mai labyardo sa e romendar le Ewropake thai le kolavrenge Barodvipurenge.
```

Output for tokenizer.py for Romani language

```
[➜  ~ head wiki.txt | python3 segmenter.py | python3 tokenizer.py

Vikipidiya
:
Vîkîpidiya
si
yek
proyekto
te
kerdivel
enchiklopedie
ande
sa
e
chibya
la
lumiyake
.
O
proyekto
shirden
andi
chon
```

Segmentation Questions

1. **How should you segment sentences with semicolons? As a single sentence or as two sentences? Should it depend on context?**

In English grammar and punctuation semicolon is used to link two independent clauses that are related in thought. Which also means that the two clauses are independent but only makes sense together and they are given equal importance. Therefore, it is best to segment sentences with semicolons as a single sentence.

For example: The house is huge; it is also old.
One sentence:
The house is huge; It is also old.
Two sentence:
The house is huge;
It is also old.
Segmenting as two sentences loses the context for the house. In NLP it is important to retain context as much as possible in most applications.

2. **Should sentences with ellipsis... be treated as a single sentence or as several sentences?**

Ellipses [...] are used in a sentence to indicate omission of words, pauses or unsaid information. It is best to treat ellipses as a single sentence as it gives meaning to the sentence that it belongs to. Segmenting it as two sentences will mean that it is treated in different contexts as ellipses are often used in quoted sentences.

3. **If there is an exclamation after the first word in the sentence should it be a separate sentence? How about if there is a comma?**

Exclamation is usually used at the end of the sentence for expressing strong feelings. If exclamation is used after the first word it is acceptable to segment as two different sentences or segment it as a single sentence. Depends on the context.
Example 1: “Help!” he shouted. “I am choking.”
After segmentation:
“Help!”
he shouted.
“I am choking.”

Example 2: Ahh! You are funny!
After segmentation:
Ahh!
You are funny!
In the first example, segmenting as multiple sentences may be a bad decision. In the second sentence segmenting as two sentences is fine.

4. **Can you think of some hard tasks for the segmenter?**

Hard parts of the segmenter is dealing with multiple languages. A particular segmenter algorithm may not work for another language.

Tokenization Questions

1. **Why should we split punctuation from the token it goes with ?**

Splitting the punctuation with the token it goes with reduces the error in text processing while performing NLP tasks.

2. **Should abbreviations with space in them be written as a single token or two tokens ?**

Abbreviations with space with them should be considered as a single token as it will lose meaning of tokenized separately.

2.1. How about numerals like 134 000 ?

Numerals should also be tokenized as a single token so that it does not lose context. Example: 134 000 soldiers died in the war.
If tokenized separately 000 does not make sense to the context.

3. **If you have a case suffix following punctuation, how should it be tokenized ?**

Example: Robert Downey Jr.
Proper nouns and case suffixes like Jr., Sr., etc. as a single token to retain the context and meaning.

4. **Should contractions and clitics be a single token or two (or more) tokens ?**

The tokenization of contraction and clitics is a hard task. It is more efficient to tokenize them separately as two tokens. Example: I’m as I am and haven't as have not.