Segmetation Practical- 2 REPORT

Ashish Patidar

For this practical, I used 2 segmenters on a wiki dump file, first segmenter is the pragmatic segementer which was covered in lecture and second segmenter is the segtok library segmenter.

Definitely, the segmenter we created in class is a basic segmenter with not many rules defined in it and the segtok segmenter is more advanced version of it.

The Pragmatice Segmenter is a basic segmenter which follows the common english rules with very little stop words included in it. Overall we cannot expect it to perform better than advanced segmenters like NLTK or Segtok.

The Segtok Segmenter is written in python which is a pattern-based segmenter, i.e. it uses regex and not Machine Learning algorithms to segment, this is also not a much advanced segmenter but is better than the pragmatic segmenter we implemented in class.
One of the common limitation of segtok segmenter is that the next sentence should either start with a Upper-Case word or a number or with a camel-cased word. Another limitation is other than human name or any name if a upper case letter is followed by a dot and some word a split would be made.

Code for Pragmatic Segmenter:

In [ ]:

```
import sys

line = sys.stdin.readline()

while line != '':
        for token in line.split(" "):
                if token.strip() == "":
                        continue
                if token[-1] in '!?':
                        sys.stdout.write(token + '\n')

                elif token[-1] == '.':
                        if token in ['etc.', 'e.g.', 'i.e.']:
                                sys.stdout.write(token + ' ')

                        else:
                                sys.stdout.write(token + '\n')
                else:
                        sys.stdout.write(token + ' ')
        line = sys.stdin.readline()
```

Output for Pragmatic Segmenter:

- The concept of "algorithm" is also used to define the notion of decidability—a notion that is central for explaining how formal systems come into being starting from a small set of axioms and rules.
- In logic, the time that an algorithm requires to complete cannot be measured, as it is not apparently related to the customary physical dimension.
- From such uncertainties, that characterize ongoing work, stems the unavailability of a definition of "algorithm" that suits both concrete (in some sense) and abstract usage of the term.

- Swift's essay is widely held to be one of the greatest examples of sustained irony in the history of the English language.
- Much of its shock value derives from the fact that the first portion of the essay describes the plight of starving beggars in Ireland, so that the reader is unprepared for the surprise of Swift's solution when he states: "A young healthy child well nursed, is, at a year old, a most delicious nourishing and wholesome food, whether stewed, roasted, baked, or boiled; and I make no doubt that it will equally serve in a fricassee, or a ragout."

- In February 2003, Armenia sent 34 peacekeepers to Kosovo where they became part of the Greek contingent.
- Officials in Yerevan have said the Armenian military plans to substantially increase the size of its peacekeeping detachment and counts on Greek assistance to the effort.
- In June 2008, Armenia sent 72 peacekeepers to Kosovo for a total of 106 peacekeepers.

- Dance sequences based on Australian football feature heavily in Robert Helpmann's 1964 ballet "The Display", his first and most famous work for the Australian Ballet.
- The game has also inspired well-known plays such as "And the Big Men Fly" (1963) by Alan Hopgood and David Williamson's "The Club" (1977), which was adapted into a 1980 film, directed by Bruce Beresford.
- Mike Brady's 1979 hit "Up There Cazaly" is considered an Australian football anthem, and references to the sport can be found in works by popular musicians, from singer-songwriter Paul Kelly to the alternative rock band TISM.
- Many Australian football video games have been released, most notably the AFL series.

- Finally, AGP allows (mandatory only in AGP 3.0) "sideband addressing", meaning that the address and data buses are separated so the address phase does not use the main address/data (AD) lines at all.
- This is done by adding an extra 8-bit "SideBand Address" bus over which the graphics controller can issue new AGP requests while other AGP data is flowing over the main 32 address/data (AD) lines.
- This results in improved overall AGP data throughput.

- Rubidium is the 16th most prevalent element in the earth's crust, however it is quite rare.
- Some minerals found in North America, South Africa, Russia, and Canada contain rubidium.

- Some potassium minerals (lepidolites, biotites, feldspar, carnallite) contain it, together with caesium.
- Pollucite, carnallite, leucite, and lepidolite are all minerals that contain rubidium.
- As a by-product of lithium extraction, it is commercially obtained from lepidolite.
- Rubidium is also found in potassium rocks and brines, which is a commercial supply.
- The majority of rubidium is now obtained as a byproduct of refining lithium.
- Rubidium is used in vacuum tubes as a getter, a material that combines with and removes trace gases from vacuum tubes.

- Albert Camus:
- This revival entailed the recruitment of clerical scholars from Mercia, Wales and abroad to enhance the tenor of the court and of the episcopacy; the establishment of a court school to educate his own children, the sons of his nobles, and intellectually promising boys of lesser birth; an attempt to require literacy in those who held offices of authority; a series of translations into the vernacular of Latin works the king deemed "most necessary for all men to know"; the compilation of a chronicle detailing the rise of Alfred's kingdom and house, with a genealogy that stretched back to Adam, thus giving the West Saxon kings a biblical ancestry.

- As part of a series of events on the bicentenary of his death, a memorial was dedicated in Westminster Abbey on 9 July 2014.
- In the service, the Dean of Westminster, Very Reverend Dr John Hall, described Phillip as follows: "This modest, yet world-class seaman, linguist, and patriot, whose selfless service laid the secure foundations on which was developed the Commonwealth of Australia, will always be remembered and honoured alongside other pioneers and inventors here in the Nave: David Livingstone, Thomas Cochrane, and Isaac Newton." A similar memorial was unveiled by the outgoing 37th Governor of New South Wales, Marie Bashir, in St James' Church, Sydney, on 31 August 2014.
- A bronze bust was installed at the Museum of Sydney, and a full-day symposium discussed his contributions to the founding of modern Australia.

- Although Cambria Iron and Steel's facilities were heavily damaged by the flood, they returned to full production within a year.
- After the flood, Carnegie built Johnstown a new library to replace the one built by Cambria's chief legal counsel Cyrus Elder, which was destroyed in the flood.
- The Carnegie-donated library is now owned by the Johnstown Area Heritage Association, and houses the Flood Museum.

Code for Segtok Segmenter:

In [ ]:

```
import sys
from segtok.segmenter import split_multi

line = sys.stdin.readline()
while line != '':
```

```
        for sent in split_multi(line):
                sys.stdout.write(str(sent) + '\n')
        line = sys.stdin.readline()
```

Output for Segtok Segmenter:

- The concept of "algorithm" is also used to define the notion of decidability—a notion that is central for explaining how formal systems come into being starting from a small set of axioms and rules.
- In logic, the time that an algorithm requires to complete cannot be measured, as it is not apparently related to the customary physical dimension.
- From such uncertainties, that characterize ongoing work, stems the unavailability of a definition of "algorithm" that suits both concrete (in some sense) and abstract usage of the term.

- Swift's essay is widely held to be one of the greatest examples of sustained irony in the history of the English language.
- Much of its shock value derives from the fact that the first portion of the essay describes the plight of starving beggars in Ireland, so that the reader is unprepared for the surprise of Swift's solution when he states: "A young healthy child well nursed, is, at a year old, a most delicious nourishing and wholesome food, whether stewed, roasted, baked, or boiled; and I make no doubt that it will equally serve in a fricassee, or a ragout."

- In February 2003, Armenia sent 34 peacekeepers to Kosovo where they became part of the Greek contingent.
- Officials in Yerevan have said the Armenian military plans to substantially increase the size of its peacekeeping detachment and counts on Greek assistance to the effort.
- In June 2008, Armenia sent 72 peacekeepers to Kosovo for a total of 106 peacekeepers.

- Dance sequences based on Australian football feature heavily in Robert Helpmann's 1964 ballet "The Display", his first and most famous work for the Australian Ballet.
- The game has also inspired well-known plays such as "And the Big Men Fly" (1963) by Alan Hopgood and David Williamson's "The Club" (1977), which was adapted into a 1980 film, directed by Bruce Beresford.
- Mike Brady's 1979 hit "Up There Cazaly" is considered an Australian football anthem, and references to the sport can be found in works by popular musicians, from singer-songwriter Paul Kelly to the alternative rock band TISM.
- Many Australian football video games have been released, most notably the AFL series.

- Finally, AGP allows (mandatory only in AGP 3.0) "sideband addressing", meaning that the address and data buses are separated so the address phase does not use the main address/data (AD) lines at all.

- This is done by adding an extra 8-bit "SideBand Address" bus over which the graphics controller can issue new AGP requests while other AGP data is flowing over the main 32 address/data (AD) lines.
- This results in improved overall AGP data throughput.

- Rubidium is the 16th most prevalent element in the earth's crust, however it is quite rare.
- Some minerals found in North America, South Africa, Russia, and Canada contain rubidium.
- Some potassium minerals (lepidolites, biotites, feldspar, carnallite) contain it, together with caesium.
- Pollucite, carnallite, leucite, and lepidolite are all minerals that contain rubidium.
- As a by-product of lithium extraction, it is commercially obtained from lepidolite.
- Rubidium is also found in potassium rocks and brines, which is a commercial supply.
- The majority of rubidium is now obtained as a byproduct of refining lithium.
- Rubidium is used in vacuum tubes as a getter, a material that combines with and removes trace gases from vacuum tubes.

- Albert Camus:
- This revival entailed the recruitment of clerical scholars from Mercia, Wales and abroad to enhance the tenor of the court and of the episcopacy; the establishment of a court school to educate his own children, the sons of his nobles, and intellectually promising boys of lesser birth; an attempt to require literacy in those who held offices of authority; a series of translations into the vernacular of Latin works the king deemed "most necessary for all men to know"; the compilation of a chronicle detailing the rise of Alfred's kingdom and house, with a genealogy that stretched back to Adam, thus giving the West Saxon kings a biblical ancestry.

- As part of a series of events on the bicentenary of his death, a memorial was dedicated in Westminster Abbey on 9 July 2014.
- In the service, the Dean of Westminster, Very Reverend Dr John Hall, described Phillip as follows: "This modest, yet world-class seaman, linguist, and patriot, whose selfless service laid the secure foundations on which was developed the Commonwealth of Australia, will always be remembered and honoured alongside other pioneers and inventors here in the Nave: David Livingstone, Thomas Cochrane, and Isaac Newton."
- A similar memorial was unveiled by the outgoing 37th Governor of New South Wales, Marie Bashir, in St James' Church, Sydney, on 31 August 2014.
- A bronze bust was installed at the Museum of Sydney, and a full-day symposium discussed his contributions to the founding of modern Australia.

- Although Cambria Iron and Steel's facilities were heavily damaged by the flood, they returned to full production within a year.
- After the flood, Carnegie built Johnstown a new library to replace the one built by Cambria's chief legal counsel Cyrus Elder, which was destroyed in the flood.

- The Carnegie-donated library is now owned by the Johnstown Area Heritage Association, and houses the Flood Museum.

Conclusion

For pragmatic segmenter 10 paragraphs are segmented into 31 lines, and for the segtok segmenter these 10 paragraphs are divided into 32 lines.
If we look at the results both the segmenters performed comparitively same, but in second last paragraph the pragmatic segmenter did not identify a sentence as full stop was followed by double-invereted commas, but was righfully identified by segtok segmenter. We can say Pragmatic segmeter is also a decent segmenter but it misses small details like this which makes other segmenters better than it.

In [ ]: