Text Mining Exercise week 4 report

Guido Zuidhof, s4160703

Timeline of Mandela's life

Date(ISO8601)	Event
1918-7-18	Nelson Mandela was born on July 18, 1918, in Mveso, Transkei, South Africa.
1918-7-18	Nelson Mandela was born Rolihlahla Mandela on July 18, 1918, in the tiny village of Mvezo, on the banks of the Mbashe River in Transkei, South Africa. "Rolihlahla" in the Xhosa language literally means "pulling the branch of a tree," but more commonly translates as "troublemaker."
1939	In 1939, Mandela enrolled at the University College of Fort Hare, the only residential center of higher learning for blacks in South Africa at the time.
1942	Becoming actively involved in the anti-apartheid movement in his 20s, Mandela joined the African National Congress in 1942.
1942	Mandela soon became actively involved in the anti-apartheid movement, joining the African National Congress in 1942.
1944	Mandela was married three times, beginning with Evelyn Ntoko Mase (m. 1944-1957).
1949	In 1949, the ANC officially adopted the Youth League's methods of boycott, strike, civil disobedience and non-cooperation, with policy goals of full citizenship, redistribution of land, trade union rights, and free and compulsory education for all children.
1952	For 20 years, Mandela directed peaceful, nonviolent acts of defiance against the South African government and its racist policies, including the 1952 Defiance Campaign and the 1955 Congress of the People.
1956	In 1956, Mandela and 150 others were arrested and charged with treason for their political advocacy (they were eventually acquitted).
1958	Mandela wed Winnie Madikizela in 1958; the couple had two daughters together, Zenani and Zindziswa, before splitting in 1996.
1959	Africanists soon broke away to form the Pan-Africanist Congress, which negatively affected the ANC; by 1959, the movement had lost much of its militant support.
1961	In 1961, Mandela, who was formerly committed to nonviolent protest, began to believe that armed struggle was the only way to achieve change.
1961	In 1961, Mandela orchestrated a three-day national workers' strike.
1963	In 1963, Mandela was brought to trial again.
1981	A 1981 memoir by South African intelligence agent Gordon Winter described a plot by the South African government to arrange for Mandela's escape so as to shoot him during the recapture; the plot was foiled by British intelligence.
1982	In 1982, Mandela and other ANC leaders were moved to Pollsmoor Prison, allegedly to enable contact between them and the South African government.
1985	In 1985, President P.W. Botha offered Mandela's release in exchange for renouncing armed struggle; the prisoner flatly rejected the offer.
1990-2-11	It wasn't until Botha suffered a stroke and was replaced by Frederik Willem de Klerk that Mandela's release was finally announced—on February 11, 1990.
1991	In 1991, Mandela was elected president of the African National Congress, with lifelong friend and colleague Oliver Tambo serving as national chairperson.
1993	In 1993, Mandela and South African President F.W. de Klerk were jointly awarded the Nobel Peace Prize for their efforts to dismantle the country's apartheid system.

Date(ISO8601)	Event
1993	In 1993, Mandela and President de Klerk were jointly awarded the Nobel Peace Prize for their work toward dismantling apartheid.
1994	Nelson Mandela became the first black president of South Africa in 1994, serving until 1999. A symbol of global peacemaking, he won the Nobel Peace Prize in 1993.
1994	In 1994, Mandela was inaugurated as South Africa's first black president.
1994	Also in 1994, Mandela published an autobiography, Long Walk to Freedom, much of which he had secretly written while in prison.
1994-4-27	And due in no small part to their work, negotiations between black and white South Africans prevailed: On April 27, 1994, South Africa held its first democratic elections.
1994-5-10	Nelson Mandela was inaugurated as the country's first black president on May 10, 1994, at the age of 77, with de Klerk as his first deputy.
1994-6	From 1994 until June 1999, Mandela worked to bring about the transition from minority rule and apartheid to black majority rule.
1995	In 1995, South Africa came to the world stage by hosting the Rugby World Cup, which brought further recognition and prestige to the young republic.
1996	In 1996, Mandela signed into law a new constitution for the nation, establishing a strong central government based on majority rule, and guaranteeing both the rights of minorities and the freedom of expression.
1999	By the 1999 general election, Nelson Mandela had retired from active politics.
2001	Mandela was diagnosed and treated for prostate cancer in 2001.
2004-6	In June 2004, at the age of 85, he announced his formal retirement from public life and returned to his native village of Qunu.
2005	In addition to advocating for peace and equality on both a national and global scale, in his later years, Mandela remained committed to the fight against AIDS—a disease that killed Mandela's son, Makgatho, in 2005.
2005	The couple had four children together: Madiba Thembekile, Makgatho (d. 2005), Makaziwe and Maki.
2007-7-18	On July 18, 2007, Mandela convened a group of world leaders, including Graca Machel (whom Mandela wed in 1998), Desmond Tutu, Kofi Annan, Ela Bhatt, Gro Harlem Brundtland, Jimmy Carter, Li Zhaoxing, Mary Robinson and Muhammad Yunus, to address some of the world's toughest issues.
2009-7-18	In 2009, Mandela's birthday (July 18) was declared "Mandela Day" to promote global peace and celebrate the South African leader's legacy.
2009-7-18	In 2009, Mandela's birthday (July 18) was declared Mandela Day, an international day to promote global peace and celebrate the South African leader's legacy.
2010	Nelson Mandela made his last public appearance at the final match of the World Cup in South Africa in 2010.
2011	He did, however, visit with U.S. first lady Michelle Obama, wife of President Barack Obama, during her trip to South Africa in 2011.
2011-1	After suffering a lung infection in January 2011, Mandela was briefly hospitalized in Johannesburg to undergo surgery for a stomach ailment in early 2012.
2012-12	Mandela would be hospitalized many times over the next several years—in December 2012, March 2013 and June 2013—for further testing and medical treatment relating to his recurrent lung infection.
2013	Two years later, Mandela married Graca Machel, with whom he remained until his death in 2013.

Date(ISO8601)	Event
2013-3	Jacob Zuma, South Africa's president, issued a statement in response to public concern over Mandela's March 2013 health scare, asking for support in the form of prayer: "We appeal to the people of South Africa and the world to pray for our beloved Madiba and his family and to keep them in their thoughts," Zuma said.
2013-6	Following his June 2013 hospital visit, Mandela's wife, Graca Machel, canceled a scheduled appearance in London to remain at her husband's his side, and his daughter, Zenani Dlamini, Argentina's South African ambassador, flew back to South Africa to be with her father.
2013-12-5	Mandela died at his home in Johannesburg on December 5, 2013, at age 95.
2013-12-5	On December 5, 2013, at the age of 95, Nelson Mandela died at his home in Johannesburg, South Africa.

Precision

Of these 45 ISO 8601 dated events, 1 is incorrect.

`1994-6 From 1994 until June 1999, Mandela worked to bring about the transition from minority rule and apartheid to black majority rule.` with date `1994-6`. The June mentioned is in 1999, not 1996.

Precision = 44/45 = 0.978. This however stretches the definition of precision. I think that, instead, precision asked here is based on the amount of date mentions it succesfully retrieved. This would be 1, as I was unable to find a missed dated event. However, some sentences contain multiple dates and years, and it mapped these only to a single timestamp.

Difficulties

This was quite challenging! The first challenge was to split the text, fortunately my exercise-2 script worked quite well. I made one small improvement to it for initials of people not being handled correctly.

The next challenge was recognizing dates in the text, I ended up using regular expressions to extract these. This worked surprisingly well.

I couldn't simply parse these and dump them into a python `datetime` object, as information would be lost whether the month and day are actually known (they have a default value), so I kept all this information alongside it. When outputting the ISO 8601 string it checks whether this was known or not.

Some dates did not have the year mentioned alongside the month and day. Here I opted to use the last-seen year, as that made the most sense because text is usually structured that way.

Finally, formatting it in a table was a challenge (which would have been a pain to do manually), so I wrote some code to write a **markdown**-style table to a file.

I had to make only one adjustment when inputting the Mandela file, that was filtering 'day' strings that were too high of a number to actually be a day of the month (the age of Mandela was recognized as the day of the month).

Source

```
import re
import codecs
import operator

from datetime import datetime


#Guido Zuidhof
#s4160703
```

```
# Text Mining, exercise 4

#From exercise 2, splits the input sentence-wise
def process_text(filename="inputnelson.txt"):

    in_file = open(filename,'r')
    full_text = in_file.read()
    in_file.close()

    #Recognize multiple empty lines, used for page numbers and footnotes
    #Place a flag there, to be added again later
    paged = re.sub(r'\n\n',r'_NEWLINE_', full_text)

    #Hyphens at the end of sentences get removed, joining the two parts
    unhyphened = paged.replace('-\n',"")

    #newlines get removed where the next character is not a capital
    collapsed = re.sub(r'\n([^A-Z])',r"\1", unhyphened)

    #Place newlines where < NOT capital, dot, a character, a capital,
    #NOT (OPTIONAL whitespace, a dot) > pattern is present
    rebuilt = re.sub(r'([^A-Z])[.][ ]([A-Z])([^?\s.])', r'\1.\n\2\3', collapsed)

    #Re-add the multiple empty lines
    rebuilt = rebuilt.replace("_NEWLINE_","\n\n")

    return rebuilt

def extract_events(text):
    sentences = text.split("\n")

    #Timeline events
    events = []
    last_year = "0000"

    for sentence in sentences:

        year = re.search(r'\d{4}', sentence)
        month = re.search(r'((Jan|Feb|Mar[^y]|Apr|May|Jun|Jul|Aug|Sep|Oct|Nov|Dec)[a-z]*)', sentence)

        day = re.search(r'[^\d{2}]((\d?\d))[^\d{2}]', sentence) if month else None

        year = year.group(0) if year else None
        month = month.group(0) if month else None
        day = day.group(0) if day and month else None

        if day: #Select only relevant part
            day = re.search(r'(\d?\d)', day).group(0)
            if int(day) > 31: #non-day, some other number
                day = None



        #Use the last seen year
        if year is not None:
            last_year = year

        #Some date was found
        if year or month:

            #For debug purposes
            #print "==\n", sentence, "\n", year, month, day

            #Construct datetime
            dt = None
            if month and day:
```

```
            if month and day:
                dt = datetime.strptime(day + " " + month + " " + last_year, '%d %B %Y')
                #print month, day, dt
            elif month:
                dt = datetime.strptime(month + " " + last_year, '%B %Y')
            else:
                dt = datetime.strptime(last_year, '%Y')

            events.append((sentence, dt, year, month, day))

    return events

#Prints ISO 8601 format in one column
#The event (text) in the other
def events_to_date_text_tuples(events):
    tups = []
    for event in events:
        text = event[0]
        dt = event[1]
        date_string = str(dt.year)
        if event[3]: #month is known
            date_string += "-"+str(dt.month)

            if event[4]: #day is known
                date_string += "-"+str(dt.day)

        tups.append((date_string,text))

    return tups

#Returns a markdown table from list of tuples
def tuples_to_table(data, header="|Date(ISO8601)|Event|"):
    table = header+"\n"
    table += "|----|---|\n"

    for x1,x2 in data:
        table += "|{0}|{1}|".format(x1,x2) +"\n"
    return table


if __name__ == "__main__":
    text = process_text()

    #Extract events from text
    events = extract_events(text)

    #Sort by datetime
    events.sort(key=operator.itemgetter(1))
    #print events

    event_text_tuples = events_to_date_text_tuples(events)
    #print event_text_tuples
    markdown_table =  tuples_to_table(event_text_tuples)

    out_file = open("table.md", "w")
    out_file.write(markdown_table)
    out_file.close()




    #print text
```