Name: Ayush Karn

Reg: 17BCE2381

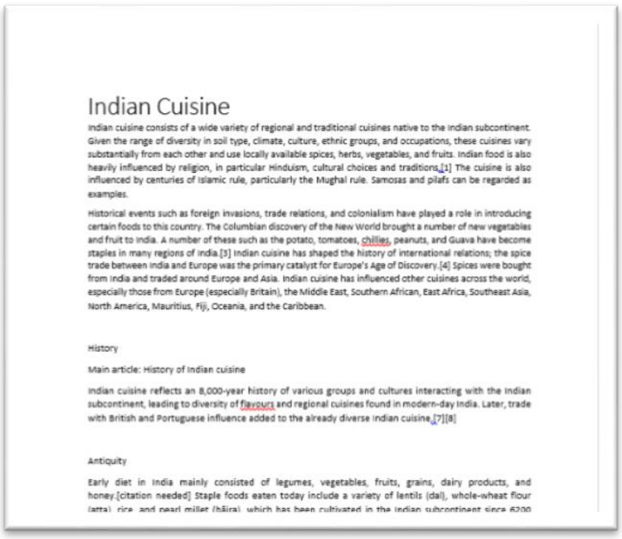
Date: 11 Jan, 2020

Magnum Opus Task II

Cosine Distance/Similarity between two documents before and after stemming

Documents:

1. Indian Cuisine.docx



Indian Cuisine

Indian cuisine consists of a wide variety of regional and traditional cuisines native to the Indian subcontinent. Given the range of diversity in soil type, climate, culture, ethnic groups, and occupations, these cuisines vary substantially from each other and use locally available spices, herbs, vegetables, and fruits. Indian food is also heavily influenced by religion, in particular Hinduism, cultural choices and traditions.[1] The cuisine is also influenced by centuries of Islamic rule, particularly the Mughal rule. Samosas and pilafs can be regarded as examples.

Historical events such as foreign invasions, trade relations, and colonialism have played a role in introducing certain foods to this country. The Columbian discovery of the New World brought a number of new vegetables and fruit to India. A number of these such as the potato, tomatoes, chillies, peanuts, and Guava have become staples in many regions of India.[3] Indian cuisine has shaped the history of international relations; the spice trade between India and Europe was the primary catalyst for Europe's Age of Discovery.[4] Spices were bought from India and traded around Europe and Asia. Indian cuisine has influenced other cuisines across the world, especially those from Europe (especially Britain), the Middle East, Southern African, East Africa, Southeast Asia, North America, Mauritius, Fiji, Oceania, and the Caribbean.

History

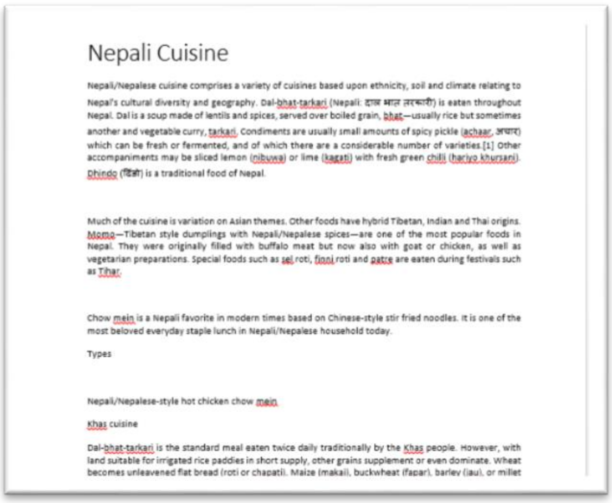
Main article: History of Indian cuisine

Indian cuisine reflects an 8,000-year history of various groups and cultures interacting with the Indian subcontinent, leading to diversity of flavours and regional cuisines found in modern-day India. Later, trade with British and Portuguese influence added to the already diverse Indian cuisine.[7][8]

Antiquity

Early diet in India mainly consisted of legumes, vegetables, fruits, grains, dairy products, and honey.[citation needed] Staple foods eaten today include a variety of lentils (dal), whole-wheat flour (atta), rice, and pearl millet (bājra), which has been cultivated in the Indian subcontinent since 6200

2. Nepali Cuisine.docx



Nepali Cuisine

Nepali/Nepalese cuisine comprises a variety of cuisines based upon ethnicity, soil and climate relating to Nepal's cultural diversity and geography. Dal-bhat-tarkari (Nepali: दाल भात तरकारी) is eaten throughout Nepal. Dal is a soup made of lentils and spices, served over boiled grain, bhat—usually rice but sometimes another and vegetable curry, tarkari. Condiments are usually small amounts of spicy pickle (achaar, अचार) which can be fresh or fermented, and of which there are a considerable number of varieties.[1] Other accompaniments may be sliced lemon (nibuwa) or lime (kagati) with fresh green chilli (hariyo khursani). Dhindo (ढिँडो) is a traditional food of Nepal.

Much of the cuisine is variation on Asian themes. Other foods have hybrid Tibetan, Indian and Thai origins. Momo—Tibetan style dumplings with Nepali/Nepalese spices—are one of the most popular foods in Nepal. They were originally filled with buffalo meat but now also with goat or chicken, as well as vegetarian preparations. Special foods such as sel roti, finni roti and patre are eaten during festivals such as Tihar.

Chow mein is a Nepali favorite in modern times based on Chinese-style stir fried noodles. It is one of the most beloved everyday staple lunch in Nepali/Nepalese household today.

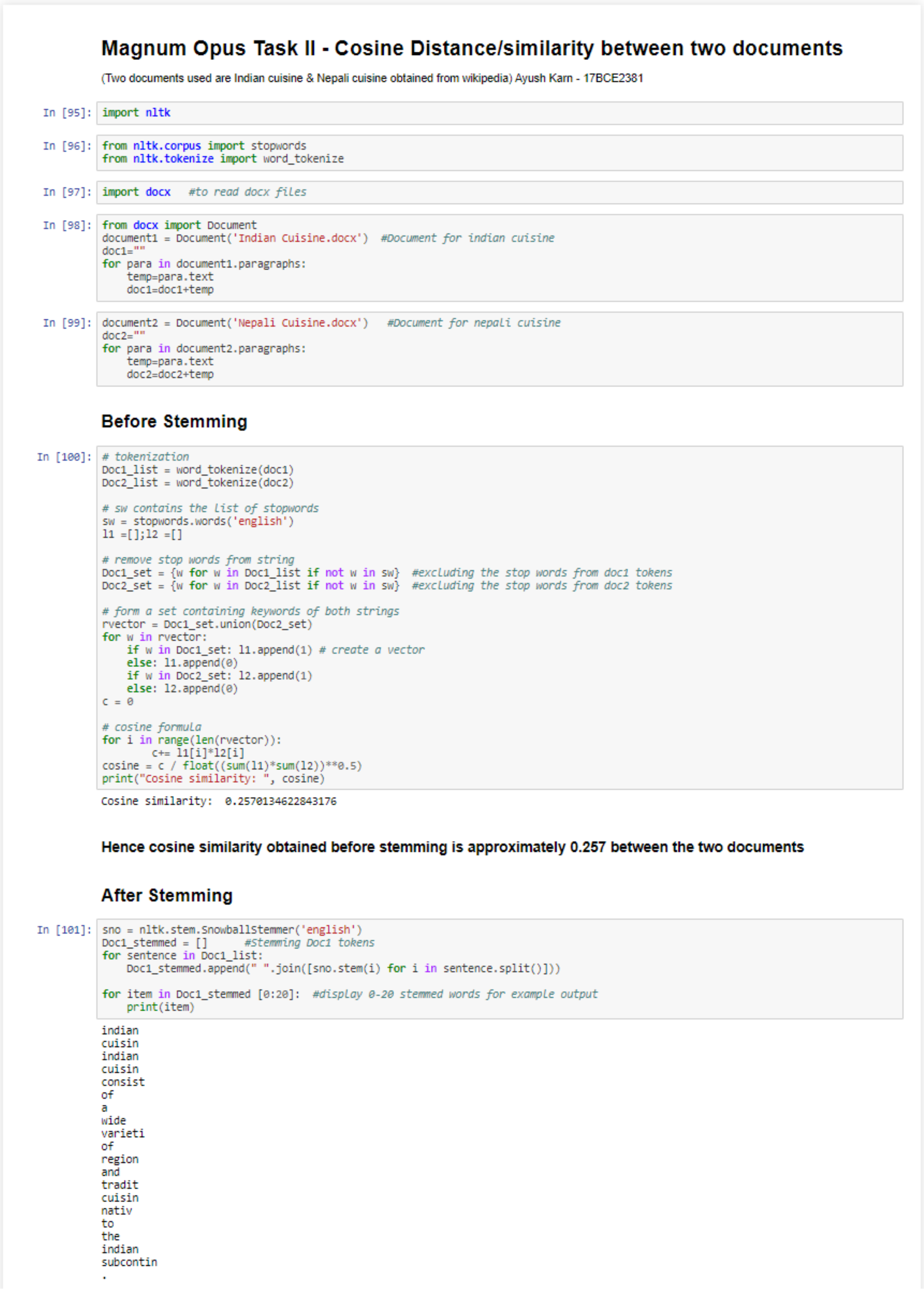
Types

Nepali/Nepalese-style hot chicken chow mein

Khas cuisine

Dal-bhat-tarkari is the standard meal eaten twice daily traditionally by the Khas people. However, with land suitable for irrigated rice paddies in short supply, other grains supplement or even dominate. Wheat becomes unleavened flat bread (roti or chapati). Maize (makai), buckwheat (fapar), barley (jau), or millet

Code Snippet:



Magnum Opus Task II - Cosine Distance/similarity between two documents

(Two documents used are Indian cuisine & Nepali cuisine obtained from wikipedia) Ayush Karn - 17BCE2381

```
In [95]: import nltk
```

```
In [96]: from nltk.corpus import stopwords
         from nltk.tokenize import word_tokenize
```

```
In [97]: import docx   #to read docx files
```

```
In [98]: from docx import Document
         document1 = Document('Indian Cuisine.docx')  #Document for indian cuisine
         doc1=""
         for para in document1.paragraphs:
             temp=para.text
             doc1=doc1+temp
```

```
In [99]: document2 = Document('Nepali Cuisine.docx')   #Document for nepali cuisine
         doc2=""
         for para in document2.paragraphs:
             temp=para.text
             doc2=doc2+temp
```

Before Stemming

```
In [100]: # tokenization
          Doc1_list = word_tokenize(doc1)
          Doc2_list = word_tokenize(doc2)

          # sw contains the list of stopwords
          sw = stopwords.words('english')
          l1 =[];l2 =[]

          # remove stop words from string
          Doc1_set = {w for w in Doc1_list if not w in sw}  #excluding the stop words from doc1 tokens
          Doc2_set = {w for w in Doc2_list if not w in sw}  #excluding the stop words from doc2 tokens

          # form a set containing keywords of both strings
          rvector = Doc1_set.union(Doc2_set)
          for w in rvector:
              if w in Doc1_set: l1.append(1) # create a vector
              else: l1.append(0)
              if w in Doc2_set: l2.append(1)
              else: l2.append(0)
          c = 0

          # cosine formula
          for i in range(len(rvector)):
                  c+= l1[i]*l2[i]
          cosine = c / float((sum(l1)*sum(l2))**0.5)
          print("Cosine similarity: ", cosine)

          Cosine similarity:  0.2570134622843176
```

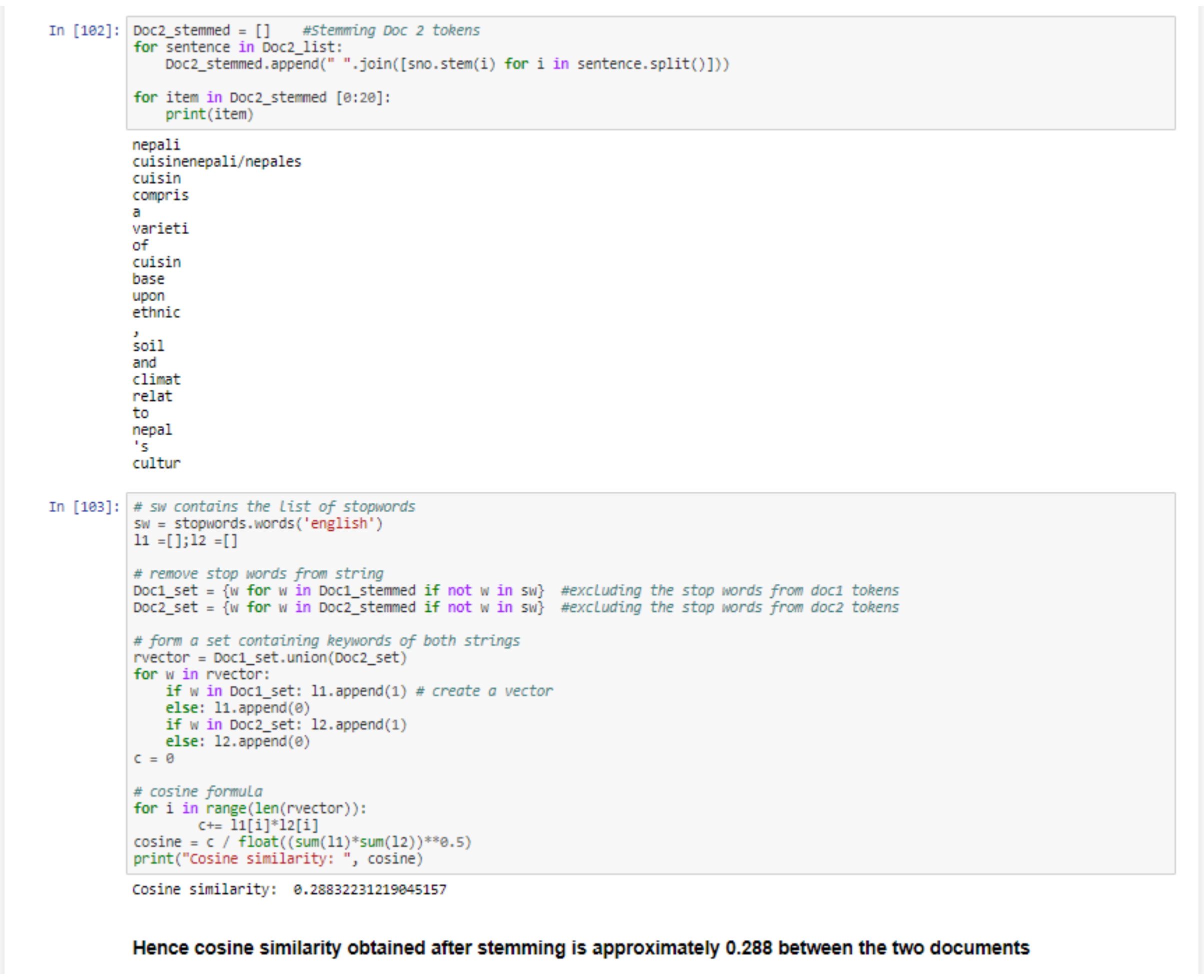
Hence cosine similarity obtained before stemming is approximately 0.257 between the two documents

After Stemming

```
In [101]: sno = nltk.stem.SnowballStemmer('english')
          Doc1_stemmed = []      #Stemming Doc1 tokens
          for sentence in Doc1_list:
              Doc1_stemmed.append(" ".join([sno.stem(i) for i in sentence.split()]))

          for item in Doc1_stemmed [0:20]:  #display 0-20 stemmed words for example output
              print(item)

          indian
          cuisin
          indian
          cuisin
          consist
          of
          a
          wide
          varieti
          of
          region
          and
          tradit
          cuisin
          nativ
          to
          the
          indian
          subcontin
          .
```



```
In [102]: Doc2_stemmed = []    #Stemming Doc 2 tokens
          for sentence in Doc2_list:
              Doc2_stemmed.append(" ".join([sno.stem(i) for i in sentence.split()]))

          for item in Doc2_stemmed [0:20]:
              print(item)
```

```
nepali
cuisinenepali/nepales
cuisin
compris
a
varieti
of
cuisin
base
upon
ethnic
,
soil
and
climat
relat
to
nepal
's
cultur
```

```
In [103]: # sw contains the list of stopwords
          sw = stopwords.words('english')
          l1 =[];l2 =[]

          # remove stop words from string
          Doc1_set = {w for w in Doc1_stemmed if not w in sw}  #excluding the stop words from doc1 tokens
          Doc2_set = {w for w in Doc2_stemmed if not w in sw}  #excluding the stop words from doc2 tokens

          # form a set containing keywords of both strings
          rvector = Doc1_set.union(Doc2_set)
          for w in rvector:
              if w in Doc1_set: l1.append(1) # create a vector
              else: l1.append(0)
              if w in Doc2_set: l2.append(1)
              else: l2.append(0)
          c = 0

          # cosine formula
          for i in range(len(rvector)):
                  c+= l1[i]*l2[i]
          cosine = c / float((sum(l1)*sum(l2))**0.5)
          print("Cosine similarity: ", cosine)
```

```
Cosine similarity:  0.28832231219045157
```

Hence cosine similarity obtained after stemming is approximately 0.288 between the two documents

Errors faced:

1. docx files weren’t readable directly. Python-docx had to be imported for reading documents while txt files can be read directly.

2. Hyperlinks (as copied from Wikipedia) were not read as a normal text by the ‘python-docx’ extension.
   So, all the hyperlinks had to be manually be removed from the docx file.

3. Inefficient Stemming algorithm, internet sources say snowball stemmer as best for general use but lots of incorrect results were obtained on stemming.

   Eg: cuisine - > cuisine
   varieties -> varieti
   relations -> relat

Conclusion

Hence, the cosine distance between two documents was found to be increased after stemming than before. With a better stemming algorithm, results would have been more accurate. Also for observance, to exactly same documents had a cosine similarity of 1.

References

- https://www.machinelearningplus.com/nlp/cosine-similarity/
- https://www.geeksforgeeks.org/python-measure-similarity-between-two-sentences-using-cosine-similarity/
- https://xapian.org/docs/stemming.html
- https://www.nltk.org/howto/stem.html