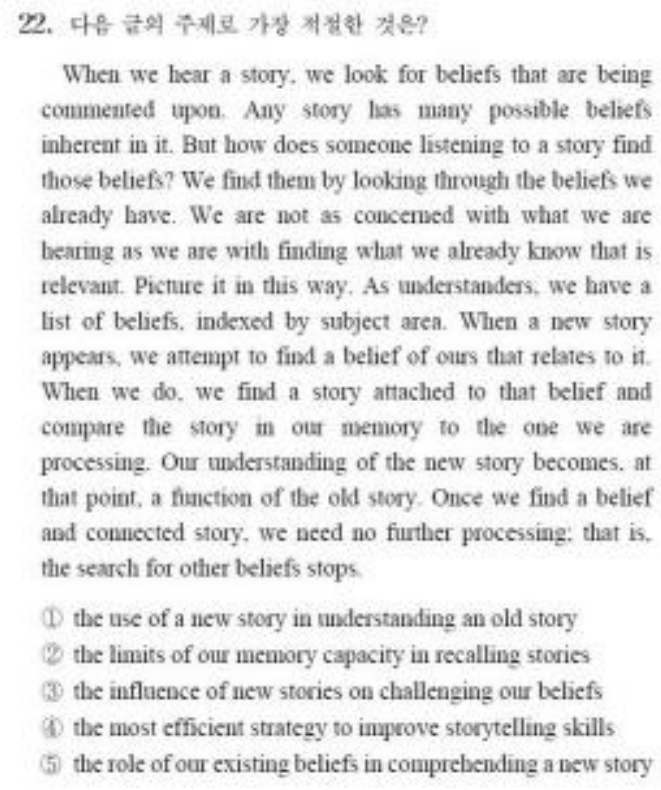
# Term project presentation
## (Korean SAT-English part topic modeling)

소프트웨어학부 20186663

전찬웅(team leader)

# Title

- The subject of the project is to guess the correct answer to the English topic(Korean SAT topic problems).



22. 다음 글의 주제로 가장 적절한 것은?

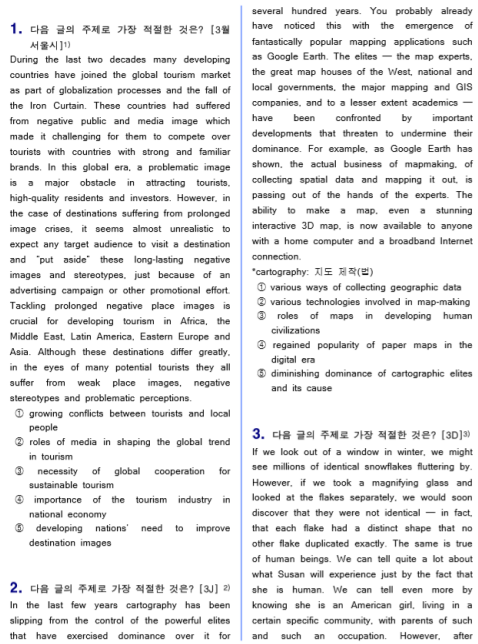
When we hear a story, we look for beliefs that are being commented upon. Any story has many possible beliefs inherent in it. But how does someone listening to a story find those beliefs? We find them by looking through the beliefs we already have. We are not as concerned with what we are hearing as we are with finding what we already know that is relevant. Picture it in this way. As understanders, we have a list of beliefs, indexed by subject area. When a new story appears, we attempt to find a belief of ours that relates to it. When we do, we find a story attached to that belief and compare the story in our memory to the one we are processing. Our understanding of the new story becomes, at that point, a function of the old story. Once we find a belief and connected story, we need no further processing; that is, the search for other beliefs stops.

① the use of a new story in understanding an old story
② the limits of our memory capacity in recalling stories
③ the influence of new stories on challenging our beliefs
④ the most efficient strategy to improve storytelling skills
⑤ the role of our existing beliefs in comprehending a new story

**To guess topic/answer**

# Data

- I used title, subject problems in Korea English SAT



**1.** 다음 글의 주제로 가장 적절한 것은? [3월 서울시][1]

During the last two decades many developing countries have joined the global tourism market as part of globalization processes and the fall of the Iron Curtain. These countries had suffered from negative public and media image which made it challenging for them to compete over tourists with countries with strong and familiar brands. In this global era, a problematic image is a major obstacle in attracting tourists, high-quality residents and investors. However, in the case of destinations suffering from prolonged image crises, it seems almost unrealistic to expect any target audience to visit a destination and "put aside" these long-lasting negative images and stereotypes, just because of an advertising campaign or other promotional effort. Tackling prolonged negative place images is crucial for developing tourism in Africa, the Middle East, Latin America, Eastern Europe and Asia. Although these destinations differ greatly, in the eyes of many potential tourists they all suffer from weak place images, negative stereotypes and problematic perceptions.

① growing conflicts between tourists and local people
② roles of media in shaping the global trend in tourism
③ necessity of global cooperation for sustainable tourism
④ importance of the tourism industry in national economy
⑤ developing nations' need to improve destination images

**2.** 다음 글의 주제로 가장 적절한 것은? [3J] [2]

In the last few years cartography has been slipping from the control of the powerful elites that have exercised dominance over it for several hundred years. You probably already have noticed this with the emergence of fantastically popular mapping applications such as Google Earth. The elites — the map experts, the great map houses of the West, national and local governments, the major mapping and GIS companies, and to a lesser extent academics — have been confronted by important developments that threaten to undermine their dominance. For example, as Google Earth has shown, the actual business of mapmaking, of collecting spatial data and mapping it out, is passing out of the hands of the experts. The ability to make a map, even a stunning interactive 3D map, is now available to anyone with a home computer and a broadband Internet connection.

*cartography: 지도 제작(법)

① various ways of collecting geographic data
② various technologies involved in map-making
③ roles of maps in developing human civilizations
④ regained popularity of paper maps in the digital era
⑤ diminishing dominance of cartographic elites and its cause

**3.** 다음 글의 주제로 가장 적절한 것은? [3D][3]

If we look out of a window in winter, we might see millions of identical snowflakes fluttering by. However, if we took a magnifying glass and looked at the flakes separately, we would soon discover that they were not identical — in fact, that each flake had a distinct shape that no other flake duplicated exactly. The same is true of human beings. We can tell quite a lot about what Susan will experience just by the fact that she is human. We can tell even more by knowing she is an American girl, living in a certain specific community, with parents of such and such an occupation. However, after

Subject problem

**1.** 다음 글의 제목으로 가장 적절한 것은? [3월 서울시][1]

With the general accessibility of photocopiers in student libraries, students tend to copy the relevant material for later use. In such cases the students are not always selective about what they copy. Often useless material is gathered that may seem important at the time but does not seem so in their study room on the night before an exam or essay due date. In addition, when most people photocopy material from books, they feel as if they have actually accomplished something. After all, a few photocopied pages in their notebook now represent information that used to be in a big, thick book. The reality of the situation is that nothing significant has been accomplished yet. The student only has the information in a transportable form. He or she has not learned anything from the material. The information content of the photocopied sheets is just as foreign as if it had been left on the library shelf.

① Information Accessibility Leads to Intellectual Advances
② Reasons You Should Keep Study Material After Exams
③ Photocopied Material: Not a Sign of Accomplishment
④ Careless Photocopying May Be Considered a Crime
⑤ Photocopier: A Contributor to Information Spread

**2.** 다음 글의 제목으로 가장 적절한 것은? [3J] [2]

As an evolutionary biologist, I am often asked whether humans are still evolving today. We certainly are. But the answer to the question of how we are changing is far more complicated. Our data suggests that the classic natural selection scenario, in which a single beneficial mutation spreads like wildfire through a population, has actually occurred relatively rarely in humans in the past 60,000 years. Rather, this mechanism of evolutionary change usually seems to require consistent environmental pressures over tens of thousands of years. Already this finding is helping to refine our understanding not only of recent human evolution but also of what our collective future might hold. For a number of the challenges currently facing our species — global climate change and many infectious diseases, for example — natural selection probably occurs too slowly to help us much.

① Effect of Social Pressure on Man
② Evolution Is Steady But Too Slow
③ Natural Selection Will Save Humans
④ How to Solve Environmental Problems
⑤ Power of Evolution to Overwhelm Culture

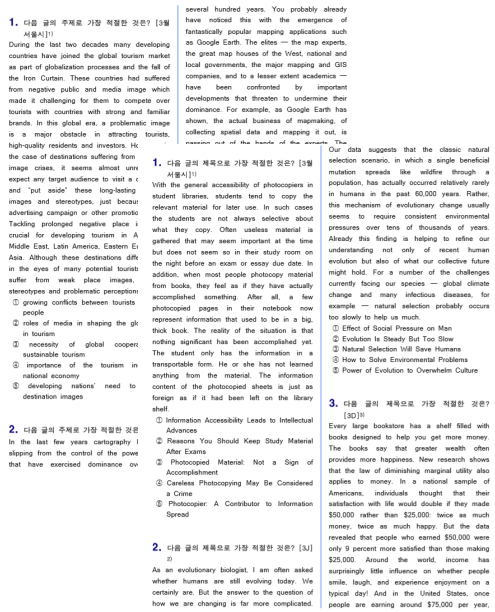
**3.** 다음 글의 제목으로 가장 적절한 것은? [3D][3]

Every large bookstore has a shelf filled with books designed to help you get more money. The books say that greater wealth often provides more happiness. New research shows that the law of diminishing marginal utility also applies to money. In a national sample of Americans, individuals thought that their satisfaction with life would double if they made $50,000 rather than $25,000: twice as much money, twice as much happy. But the data revealed that people who earned $50,000 were only 9 percent more satisfied than those making $25,000. Around the world, income has surprisingly little influence on whether people smile, laugh, and experience enjoyment on a typical day! And in the United States, once people are earning around $75,000 per year,

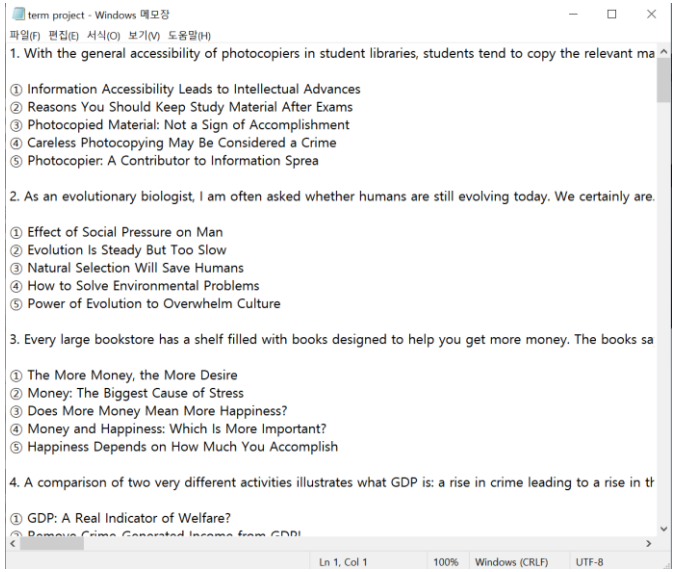
Title problem

# Data preprocessing

- I used title, subject problems in Korea English SAT
    1. first all, I processed .pdf data to .txt data



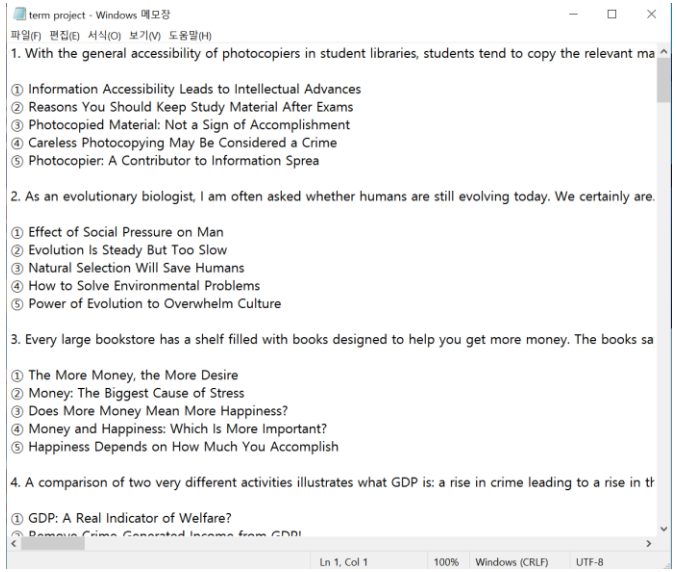
.pdf data

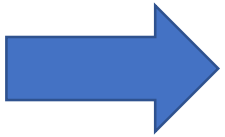


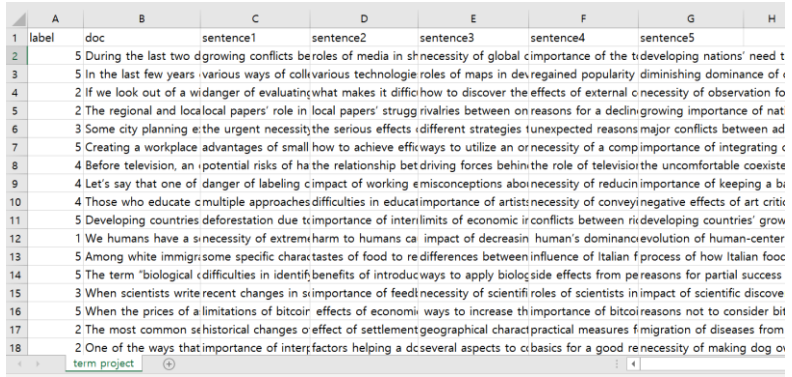

.txt data

# Data preprocessing

- I used title, subject problems in Korea English SAT
  2. And then, I processed .txt data to .csv data





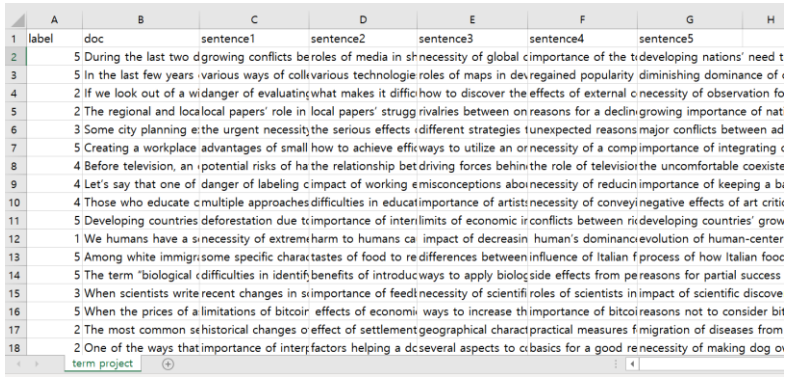

.txt data

.csv data

# Data preprocessing

- I used title, subject problems in Korea English SAT

  3. Finally, I used .csv data in pandas for data analysis



| | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| 1 | label | doc | sentence1 | sentence2 | sentence3 | sentence4 | sentence5 | |
| 2 | 5 | During the last two d | growing conflicts be | roles of media in sh | necessity of global c | importance of the t | developing nations' need t | |
| 3 | 5 | In the last few years | various ways of colle | various technologie | roles of maps in dev | regained popularity | diminishing dominance of | |
| 4 | 2 | If we look out of a wi | danger of evaluating | what makes it diffic | how to discover the | effects of external c | necessity of observation fo | |
| 5 | 2 | The regional and loca | local papers' role in | local papers' strugg | rivalries between on | reasons for a declin | growing importance of nat | |
| 6 | 3 | Some city planning e | the urgent necessity | the serious effects | different strategies t | unexpected reasons | major conflicts between ad | |
| 7 | 5 | Creating a workplace | advantages of small | how to achieve effic | ways to utilize an or | necessity of a comp | importance of integrating | |
| 8 | 4 | Before television, an | potential risks of ha | the relationship bet | driving forces behin | the role of television | the uncomfortable coexiste | |
| 9 | 4 | Let's say that one of | danger of labeling c | impact of working e | misconceptions abo | necessity of reducin | importance of keeping a ba | |
| 10 | 4 | Those who educate c | multiple approaches | difficulties in educa | importance of artist | necessity of conveyi | negative effects of art criti | |
| 11 | 5 | Developing countries | deforestation due to | importance of interi | limits of economic ir | conflicts between ric | developing countries' grow | |
| 12 | 1 | We humans have a s | necessity of extreme | harm to humans ca | impact of decreasin | human's dominance | evolution of human-center | |
| 13 | 5 | Among white immigr | some specific charac | tastes of food to re | differences between | influence of Italian f | process of how Italian fooc | |
| 14 | 5 | The term "biological c | difficulties in identify | benefits of introduc | ways to apply biolog | side effects from pe | reasons for partial success | |
| 15 | 3 | When scientists write | recent changes in s | importance of feedl | necessity of scientifi | roles of scientists in | impact of scientific discove | |
| 16 | 5 | When the prices of a | limitations of bitcoir | effects of economi | ways to increase th | importance of bitcoi | reasons not to consider bit | |
| 17 | 2 | The most common se | historical changes o | effect of settlement | geographical charact | practical measures f | migration of diseases from | |
| 18 | 2 | One of the ways that | importance of interp | factors helping a dc | several aspects to cc | basics for a good re | necessity of making dog o | |

term project

.csv data

```
In [7]: import pandas as pd
        df = pd.read_csv("term project.csv")
        df[:]
```

Out[7]:

| | label | doc | sentence1 | sentence2 | sentence3 | sentence4 | sentence5 |
|---|---|---|---|---|---|---|---|
| 0 | 5 | During the last two decades many developing co... | growing conflicts between tourists and local p... | roles of media in shaping the global trend in ... | necessity of global cooperation for sustainabl... | importance of the tourism industry in national... | developing nations' need to improve destinatio... |
| 1 | 5 | In the last few years cartography has been sli... | various ways of collecting geographic data | various technologies involved in map-making | roles of maps in developing human civilizations | regained popularity of paper maps in the digit... | diminishing dominance of cartographic elites a... |
| 2 | 2 | If we look out of a window in winter, we might... | danger of evaluating things through appearance | what makes it difficult to predict a human's life | how to discover the beauty of natural elements | effects of external conditions on one's person... | necessity of observation for understanding obj... |
| 3 | 2 | The regional and local press faces its own dis... | local papers' role in providing community news | local papers' struggles in securing revenue on... | rivalries between online news agencies and pap... | reasons for a decline in the quality of online... | growing importance of national news over local... |
| 4 | 3 | Some city planning experts called for legislat... | the urgent necessity of regular safety drills ... | the serious effects of tech-addiction on cogni... | different strategies to address the problem of... | unexpected reasons why legislation against tex... | major conflicts between advanced technology an... |
| ... | ... | ... | ... | ... | ... | ... | ... |
| 62 | 1 | It is difficult to understand what it means to... | Good Stories Put Flesh on Abstract Ideas | The Fine Line Between Fact and Fiction | There's No Better Time Killer Than a Novel | No More Fiction, Now It's Time for Nonfiction! | Don't Read a Summary First, Just Read It Through! |
| 63 | 1 | Every growing male needs to feel that he bring... | Being Helpful to Others: A Basis for Male Iden... | Is Our Society Ruled by a Male-Dominated Culture? | Male and Female Brains: Similarities and Diffe... | The Influence of Social Media Use on Gender Id... | Is Gender Identity Determined Biologically or ... |
| 64 | 5 | We speak of the complex network of meanings of... | Why Do Writers Try to Assign a Meaning to Life? | Imagination: The Most Important Quality of a W... | There Are Still Writers Who Write with Pen and... | The Moment of Writing: Not an Escape, But an A... | Do the Meanings of a Work Lie Just Within the ... |
| 65 | 1 | Within every problem, difficulty, or hardship ... | Making Good Things Come out of Adversity | Not Easy to Distinguish Between Good and Evil | You Don't Have to Respond to Every Opposition | Avoid People Who Will Rob You of Your Happiness | Don't Turn Your Past Troubles into Your Presen... |
| 66 | 1 | A defining element of catastrophes is the magn... | Insensitivity to Mass Tragedy: We Are Lost in ... | Power of Numbers: A Way of Classifying Natural... | How to Reach Out a Hand to People in Desperate... | Preventing Potential Losses Through Technology | Be Careful, Numbers Magnify Feelings! |

67 rows × 7 columns

Data by pandas

# Vectorization

- I used Count Vectorization and TF-IDF Vectorization
  - Count Vectorizer : Simply create a vector with the frequency of words used.
  - TF IDF Vectorizer : Weight is given by comparing the number of words used in the document with the number of words used in the entire document.

$$w_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right)$$

**TF-IDF**

Term $x$ within document $y$

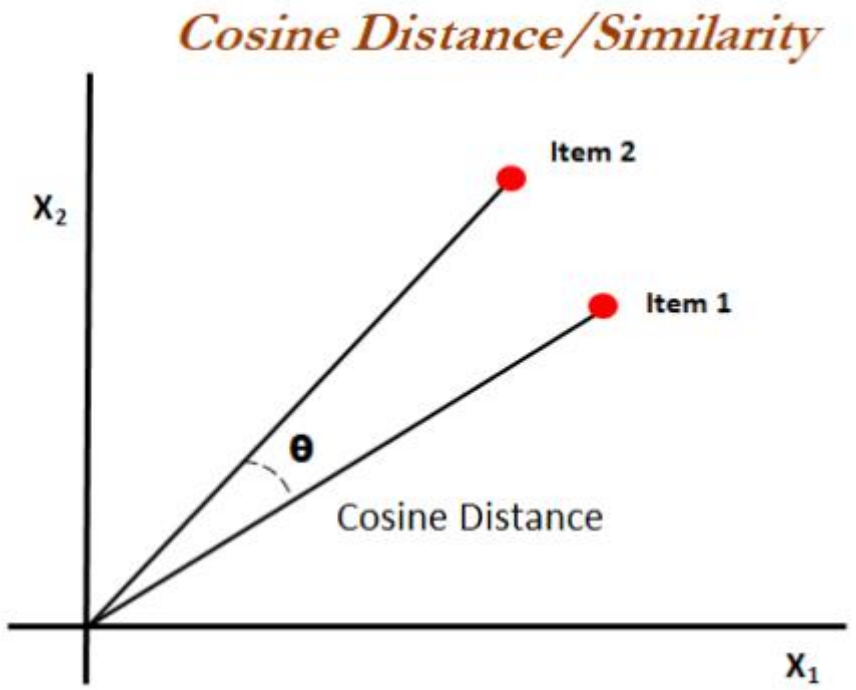
$tf_{x,y}$ = frequency of $x$ in $y$

$df_x$ = number of documents containing $x$

$N$ = total number of documents

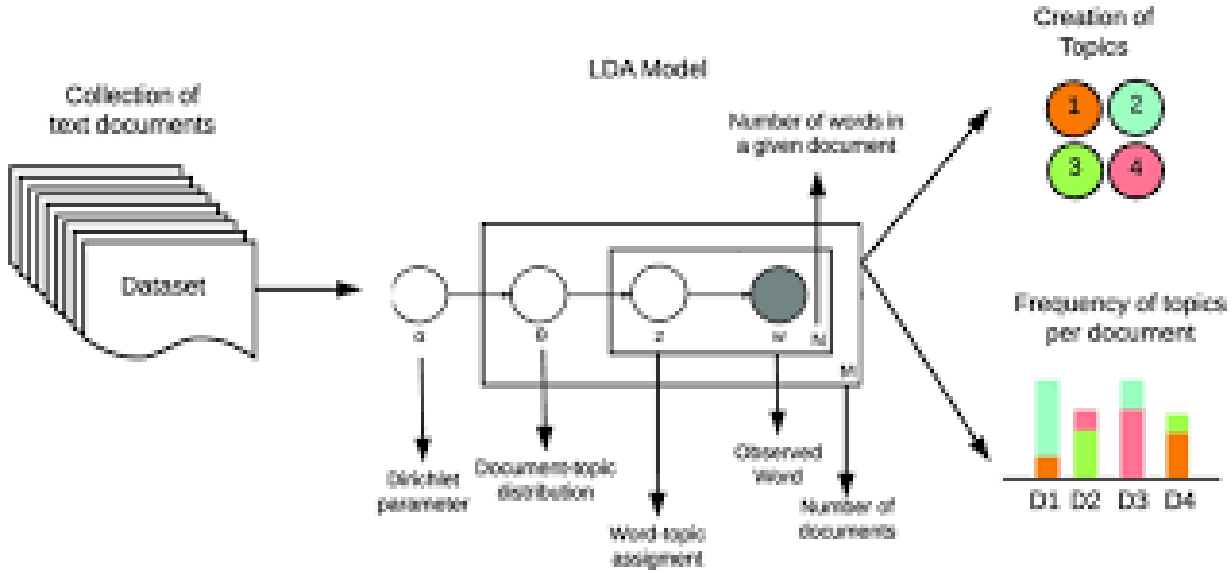
# Cosine Similarity

- I used cosine similarity to measure the similarity between two vectorized documents.



# Latent Dirichlet Allocation

- Latent Dirichlet allocation(LDA) is one of the probabilistic topic model techniques for describing what topics exist in each document for a given document.
- In this project, I used LDA to extract words related to the topic.



# TF-IDF Vectorization

- TF-IDF Vectorizer → cosine similarity

**TF-IDF Vectorization**

vectorizer

Cosine similarity

```
my_label = []
for i in range(len(df_doc)):
    dfdf = df_doc.iloc[i]
    tfidf_vect = TfidfVectorizer(stop_words='english',ngram_range=(1,2), max_df=100)
    dd = tfidf_vect.fit_transform(dfdf)
    dd = dd.todense()
    a = 0
    a_num = 0
    vect0 = np.array(dd[0]).reshape(-1,)
    for j in range(1,6):
        vect1 = np.array(dd[j]).reshape(-1,)
        b = cos_similarity(vect0,vect1)
        if b > a:
            a = b
            a_num = j
    my_label.append(a_num)
print('accuarcy',metrics.accuracy_score(df_label,my_label))
```

accuracy

```
accuarcy 0.47761194029850745
```

# Count Vectorization

- Count Vectorizer → cosine similarity

**Count Vectorization**

vectorizer

Cosine similarity

```
my_label = []
for i in range(len(df_doc)):
    dfdf = df_doc.iloc[i]
    tfidf_vect = CountVectorizer(stop_words='english',ngram_range=(1,2), max_df=100, min_df=1)
    dd = tfidf_vect.fit_transform(dfdf)
    dd = dd.todense()
    a = 0
    a_num = 0
    vect0 = np.array(dd[0]).reshape(-1,)
    for j in range(1,6):
        vect1 = np.array(dd[j]).reshape(-1,)
        b = cos_similarity(vect0,vect1)
        if b > a:
            a = b
            a_num = j
    my_label.append(a_num)
print('accuarcy',metrics.accuracy_score(df_label,my_label))
```

accuracy

```
accuarcy 0.4925373134328358
```

# TF-IDF Vectorization & LDA

- TF-IDF Vectorizer → LDA → cosine similarity

**TF-IDF Vectorization & Latent Dirichlet Allocation**

vectorizer

LDA

Cosine similarity

```
from sklearn.decomposition import LatentDirichletAllocation

my_label = []
for i in range(len(df_doc)):
    dfdf = df_doc.iloc[i]
    tfidf_vect = TfidfVectorizer(stop_words='english',ngram_range=(1,2), max_df=100)
    dd = tfidf_vect.fit_transform(dfdf)
    dd = dd.T
    lda = LatentDirichletAllocation(n_components=30,random_state=0)
    lda.fit(dd)
    dd = lda.components_.T
    a = 0
    a_num = 0
    vect0 = np.array(dd[0]).reshape(-1,)
    for j in range(1,6):
        vect1 = np.array(dd[j]).reshape(-1,)
        b = cos_similarity(vect0,vect1)
        if b > a:
            a = b
            a_num = j
    my_label.append(a_num)
print('accuarcy',metrics.accuracy_score(df_label,my_label))
```

accuracy

```
accuarcy 0.3880597014925373
```

# Count Vectorization & LDA

- Count Vectorizer → LDA → cosine similarity

**Count Vectorization & Latent Dirichlet Allocation**

vectorizer

LDA

Cosine similarity

```
from sklearn.decomposition import LatentDirichletAllocation

my_label = []
for i in range(len(df_doc)):
    dfdf = df_doc.iloc[i]
    tfidf_vect = CountVectorizer(stop_words='english',ngram_range=(1,2), max_df=100)
    dd = tfidf_vect.fit_transform(dfdf)
    dd = dd.T
    lda = LatentDirichletAllocation(n_components=10,random_state=0)
    lda.fit(dd)
    dd = lda.components_.T
    a = 0
    a_num = 0
    vect0 = np.array(dd[0]).reshape(-1,)
    for j in range(1,6):
        vect1 = np.array(dd[j]).reshape(-1,)
        b = cos_similarity(vect0,vect1)
        if b > a:
            a = b
            a_num = j
    my_label.append(a_num)
print('accuarcy',metrics.accuracy_score(df_label,my_label))
```

accuracy

```
accuarcy 0.26865671641791045
```

# Conclusion

- If using this programming, we can get about 50 points in the English test.
- It was difficult to find a meaningful one in each sentence because the amount of each sentence was too small length compared to the document. Therefore, the value of count vectorization, which simply vectorizes into the number of words, came out the highest.
- In short sentences, the meaning of the sentence being LDA was difficult to find, so the process of LDA was hindered.

# Thanks