NLTK

January 25, 2024

1 Assignment 1 - NLTK

Given a collection of documents, conduct text preprocessing including tokenization, stop words removal, stemming, tf-idf calculation, and pairwise cosine similarity calculation using NLTK.

Matthew Acs

1.1 Part 1

Install Python and NLTK

NLTK and python can be configured on a cloud based Jupyter notebook such as Google Colab.

The code below installs NLTK and related dependencies.

[ ]:
```
import nltk
from nltk.tokenize import word_tokenize
from nltk.corpus import stopwords
from nltk.stem import PorterStemmer
from nltk.tokenize import word_tokenize

nltk.download('punkt')
nltk.download('stopwords')
```

```
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Unzipping tokenizers/punkt.zip.
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Unzipping corpora/stopwords.zip.
```

[ ]:
```
True
```

The code below downloads the text data from a GitHub repository and prints it.

[ ]:
```
import requests
import textwrap

url = 'https://raw.githubusercontent.com/matthewaaa123/CAP-6776-6640/main/text_1.
 ↪txt'
text_1 = requests.get(url).text

url = 'https://raw.githubusercontent.com/matthewaaa123/CAP-6776-6640/main/text_2.
 ↪txt'
text_2 = requests.get(url).text

url = 'https://raw.githubusercontent.com/matthewaaa123/CAP-6776-6640/main/text_3.
 ↪txt'
text_3 = requests.get(url).text

print("Text 1")
print("------")
print(textwrap.fill(text_1, width=100))
print()

print("Text 2")
print("------")
print(textwrap.fill(text_2, width=100))
print()

print("Text 3")
print("------")
print(textwrap.fill(text_3, width=100))
```

```
Text 1
------
Channel tunnel operator Eurotunnel on Monday announced details of a deal giving
bank creditors 45.5
percent of the company in return for wiping out 1.0 billion pounds ($1.6
billion) of its massive
debts. The long-awaited but highly complex restructuring of nearly nearly nine
billion pounds of
debt and unpaid interest throws the company a lifeline which could secure what
is still likely to be
```

a difficult future. The deal, announced simultaneously in Paris and London, brings the company back
from the brink of bankruptcy but leaves current shareholders, who have already seen their investment
dwindle, owning only 54.5 percent of the company. "We have fixed and capped the interest payments
and arranged only to pay what is available in cash," Eurotunnel co-chairman Alastair Morton told
reporters at a news conference. "Avoiding having to do this again is the name of the game." Morton
said the plan provides the Anglo-French company with the medium term financial stability to
consolidate its commercial position and develop its operations, adding that the firm was now making
a profit before interest. Although shareholders will see their holdings diluted, they were offered
the prospect of a brighter future and urged to be patient after months of uncertainty while
Eurotunnel wrestled to reduce the crippling interest payments negotiated during the tunnel's
construction. Eurotunnel, which has taken around half of the market in the busiest cross-Channel
route from the European ferry companies, said a strong operating performance could allow it to pay
its first dividend within the next 10 years. French co-chairman Patrick Ponsolle told reporters at a
Paris news conference that the dividend could come as early as 2004 if the company performed "very
well". Eurotunnel and the banks have come up with an ingenious formula to help the company get over
the early years of the deal when, despite the swaps of debt for equity and bonds, it will still not
be able to afford the annual interest bill of 400 million pounds. If its revenue, after costs and
depreciation, is less than 400 million pounds, then the company will issue "Stabilisation notes" to
a maximum of 1.85 billion pounds to the banks. Eurotunnel would not pay interest on these notes
(which would constitute a debt issue) for ten years. Analysts said that under the deal, Eurotunnel's
ability to finance its debt would become sustainable, at least for a few years. "If you look at the
current cash flow of between 150 and 200 million pounds a year, what they can't find (to meet the
bill) they will roll forward into the stabilisation notes, and they can keep that going for seven,
eight, nine years," said an analyst at one major investment bank. "So they are here for that time,"

he added. The company said in a statement there was still considerable work to be done to finalise
and agree the details of the plan before it can be submitted to shareholders and the bank group for
approval, probably early in the Spring of 1997. Eurotunnel said the debt-for-equity swap would be at
130 pence, or 10.40 francs, per share -- considerably below the level of 160 pence widely reported
in the run up to the deal The company said a further 3.7 billion pounds of debt would be converted
into new financial instruments and existing shareholders would be able to participate in this issue.
If they choose not to take up free warrants entitling them to subscribe to this, Eurotunnel said
shareholders' interests may be reduced further to just over 39 percent of the company by the end of
December 2003. Eurotunnel's shares, which were suspended last week at 113.5 pence ahead of Monday's
announcement, will resume trading on Tuesday. Shareholders and all 225 creditor banks have to agree
the deal. "I'm hopeful but I'm not taking it (approval) for granted," Morton admitted, "Shareholders
are pretty angry in France." Asked what would happen if the banks reject the deal, Morton said,
"Nobody wants a collapse, nobody wants a doomsday scenario." ($1=.6393 Pound)

Text 2

Anglo-French Channel Tunnel operator Eurotunnel Monday announced a deal giving its creditor banks
45.5 percent of the company in return for wiping out one billion pounds ($1.56 billion) of its debt.
The long-awaited restructuring brings to an end months of wrangling between Eurotunnel and the 225
banks to which it owes nearly nine billion pounds ($14.1 billion). The deal, announced
simultaneously in Paris and London, brings the company back from the brink of insolvency but leaves
shareholders owning only 54.5 percent of the company. "The restructuring plan provides Eurotunnel
with the medium-term financial stability to allow it to consolidate its substantial commercial
achievements to date and to develop its operations," Eurotunnel co- chairman Alastair Morton said.
The firm was now making a profit before interest, he added. Although shareholders will see their
interests diluted, they were offered the prospect of a brighter future after months of uncertainty

while Eurotunnel wrestled to reduce crippling interest payments negotiated during the tunnel's
construction. Eurotunnel, which has taken around half the cross-Channel market from the European
ferry companies, said a strong operating performance could allow it to pay its first dividend within
the next 10 years. French co-chairman Patrick Ponsolle said shareholders would have to be patient
before they could reap the benefits of the company's success. He called the debt restructuring plan
"an acceptable compromise" for holders of Eurotunnel shares. The company said there was still
considerable work to be done to finalise and agree on the details of the plan before it can be
submitted to shareholders and the full 225 bank syndicate for approval, probably early in 1997.
Monday's announcement followed two weeks of highly secretive negotiations between Eurotunnel and its
six leading banks. This was extended to the 24 "instructing banks" at a meeting late last week in
London. Eurotunnel said the debt-for-equity swap would be at 130 pence, or 10.40 francs, per share.
That is considerably below the level of around 160 pence widely reported before announcement of the
deal, and will reduce outstanding debt of 8.7 billion pounds ($13.6 billion) by 1.0 billion ($1.56
billion). The company said a further 3.7 billion pounds ($5.8 billion) of debt would be converted
into new financial instruments and existing shareholders would be able to participate in this issue.
If they choose not to take up free warrants entitling them to subscribe to this, Eurotunnel said
shareholders' interests may be reduced further to just over 39 percent of the company by the end of
December 2003. Eurotunnel's shares, which were suspended last week at 113.5 pence ahead of Monday's
announcement, should resume trading on Tuesday, the company said.

Text 3

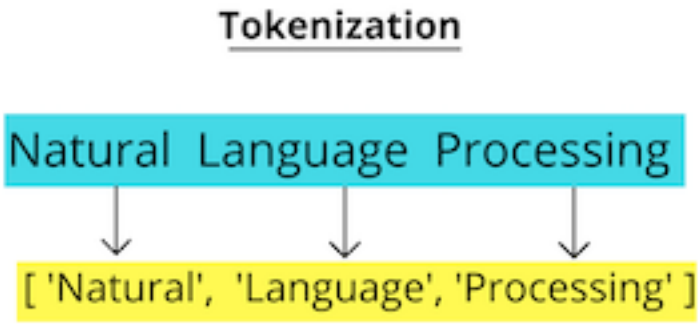
Anglo-French Channel Tunnel operator Eurotunnel on Monday announced a deal giving creditor banks
45.5 percent of the company in return for wiping out one billion pounds ($1.56 billion) of its debt
mountain. The long-awaited restructuring brings to an end months of wrangling between Eurotunnel and
the 225 banks to which it owes nearly nine billion pounds. The deal, announced simultaneously in

Paris and London, brings the company back from the brink of insolvency but leaves shareholders
owning only 54.5 percent of the company. "The restructuring plan provides Eurotunnel with the medium
term financial stability to allow it to consolidate its substantial commercial achievements to date
and to develop its operations," Eurotunnel co- chairman Alastair Morton said. The firm was now
making a profit before interest, he added. Although shareholders will see their interests diluted,
they were offered the prospect of a brighter future after months of uncertainty while Eurotunnel
wrestled to reduce crippling interest payments negotiated during the tunnel's construction.
Eurotunnel, which has taken around half the cross-Channel market from the European ferry companies,
said a strong operating performance could allow it to pay its first dividend within the next 10
years. French co-chairman Patrick Ponsolle said shareholders would have to be patient before they
could reap the benefits of the company's success. He called the debt restructuring plan "an
acceptable compromise" for holders of Eurotunnel shares. The company said in a statement there was
still considerable work to be done to finalise and agree the details of the plan before it can be
submitted to shareholders and the full 225 bank syndicate for approval, probably early in 1997.
Monday's announcement followed two weeks of highly secretive negotiations between Eurotunnel and its
six leading banks. This was extended to the 24 "instructing banks" at a meeting late last week in
London. Eurotunnel said the debt-for-equity swap would be at 130 pence, or 10.40 francs, per share.
That is considerably below the level of around 160 pence widely reported in the run up to the deal,
and will reduce outstanding debt of 8.7 billion pounds by 1.0 billion. The company said a further
3.7 billion pounds of debt would be converted into new financial instruments and existing
shareholders would be able to participate in this issue. If they choose not to take up free warrants
entitling them to subscribe to this, Eurotunnel said shareholders' interests may be reduced further
to just over 39 percent of the company by the end of December 2003. Eurotunnel's shares, which were
suspended last week at 113.5 pence ahead of Monday's announcement, should resume trading on Tuesday,

```
the company said. ($1=.6393 Pound)
```

1.2 Part 2

Tokenize the documents into words, remove stop words, and conduct stemming



Preprocessing is an important step before tf-idf vectors can be computed.

The code below tokenizes the text documents into words, removes stop words, and then applies stemming using the porter stemming algorithm.

[ ]:
```
ps = PorterStemmer()

stop_words = set(stopwords.words("english"))

text_1_tokenized = word_tokenize(text_1)
text_1_processed = []

text_2_tokenized = word_tokenize(text_2)
text_2_processed = []

text_3_tokenized = word_tokenize(text_3)
text_3_processed = []

for w in text_1_tokenized:
  if w not in stop_words:
    text_1_processed.append(ps.stem(w))

for w in text_2_tokenized:
  if w not in stop_words:
    text_2_processed.append(ps.stem(w))

for w in text_3_tokenized:
  if w not in stop_words:
    text_3_processed.append(ps.stem(w))
```

The code below prints the tokens of each document.

[ ]:
```
print("Text 1 Tokens")
print("----------------")
print(text_1_processed)
print()

print("Text 2 Tokens")
print("----------------")
print(text_2_processed)
print()

print("Text 3 Tokens")
print("----------------")
print(text_3_processed)
print()
```

```
Text 1 Tokens
----------------
['channel', 'tunnel', 'oper', 'eurotunnel', 'monday', 'announc', 'detail',
'deal', 'give', 'bank', 'creditor', '45.5', 'percent', 'compani', 'return',
'wipe', '1.0', 'billion', 'pound', '(', '$', '1.6', 'billion', ')', 'massiv',
'debt', '.', 'the', 'long-await', 'highli', 'complex', 'restructur', 'nearli',
'nearli', 'nine', 'billion', 'pound', 'debt', 'unpaid', 'interest', 'throw',
'compani', 'lifelin', 'could', 'secur', 'still', 'like', 'difficult', 'futur',
'.', 'the', 'deal', ',', 'announc', 'simultan', 'pari', 'london', ',', 'bring',
'compani', 'back', 'brink', 'bankruptci', 'leav', 'current', 'sharehold', ',',
'alreadi', 'seen', 'invest', 'dwindl', ',', 'own', '54.5', 'percent', 'compani',
'.', '``', 'we', 'fix', 'cap', 'interest', 'payment', 'arrang', 'pay', 'avail',
'cash', ',', "''", 'eurotunnel', 'co-chairman', 'alastair', 'morton', 'told',
'report', 'news', 'confer', '.', '``', 'avoid', 'name', 'game', '.', "''",
'morton', 'said', 'plan', 'provid', 'anglo-french', 'compani', 'medium', 'term',
'financi', 'stabil', 'consolid', 'commerci', 'posit', 'develop', 'oper', ',',
'ad', 'firm', 'make', 'profit', 'interest', '.', 'although', 'sharehold', 'see',
'hold', 'dilut', ',', 'offer', 'prospect', 'brighter', 'futur', 'urg',
'patient', 'month', 'uncertainti', 'eurotunnel', 'wrestl', 'reduc', 'crippl',
'interest', 'payment', 'negoti', 'tunnel', "'s", 'construct', '.', 'eurotunnel',
',', 'taken', 'around', 'half', 'market', 'busiest', 'cross-channel', 'rout',
'european', 'ferri', 'compani', ',', 'said', 'strong', 'oper', 'perform',
'could', 'allow', 'pay', 'first', 'dividend', 'within', 'next', '10', 'year',
'.', 'french', 'co-chairman', 'patrick', 'ponsol', 'told', 'report', 'pari',
'news', 'confer', 'dividend', 'could', 'come', 'earli', '2004', 'compani',
'perform', '``', 'well', "''", '.', 'eurotunnel', 'bank', 'come', 'ingeni',
'formula', 'help', 'compani', 'get', 'earli', 'year', 'deal', ',', 'despit',
'swap', 'debt', 'equiti', 'bond', ',', 'still', 'abl', 'afford', 'annual',
'interest', 'bill', '400', 'million', 'pound', '.', 'if', 'revenu', ',', 'cost',
'depreci', ',', 'less', '400', 'million', 'pound', ',', 'compani', 'issu', '``',
'stabilis', 'note', "''", 'maximum', '1.85', 'billion', 'pound', 'bank', '.',
'eurotunnel', 'would', 'pay', 'interest', 'note', '(', 'would', 'constitut',
'debt', 'issu', ')', 'ten', 'year', '.', 'analyst', 'said', 'deal', ',',
```

```
'eurotunnel', "'s", 'abil', 'financ', 'debt', 'would', 'becom', 'sustain', ',',
'least', 'year', '.', '``', 'if', 'look', 'current', 'cash', 'flow', '150',
'200', 'million', 'pound', 'year', ',', 'ca', "n't", 'find', '(', 'meet',
'bill', ')', 'roll', 'forward', 'stabilis', 'note', ',', 'keep', 'go', 'seven',
',', 'eight', ',', 'nine', 'year', ',', "''", 'said', 'analyst', 'one', 'major',
'invest', 'bank', '.', '``', 'so', 'time', ',', "''", 'ad', '.', 'the',
'compani', 'said', 'statement', 'still', 'consider', 'work', 'done', 'finalis',
'agre', 'detail', 'plan', 'submit', 'sharehold', 'bank', 'group', 'approv', ',',
'probabl', 'earli', 'spring', '1997', '.', 'eurotunnel', 'said', 'debt-for-equ',
'swap', 'would', '130', 'penc', ',', '10.40', 'franc', ',', 'per', 'share',
'--', 'consider', 'level', '160', 'penc', 'wide', 'report', 'run', 'deal',
'the', 'compani', 'said', '3.7', 'billion', 'pound', 'debt', 'would', 'convert',
'new', 'financi', 'instrument', 'exist', 'sharehold', 'would', 'abl',
'particip', 'issu', '.', 'if', 'choos', 'take', 'free', 'warrant', 'entitl',
'subscrib', ',', 'eurotunnel', 'said', 'sharehold', "''", 'interest', 'may',
'reduc', '39', 'percent', 'compani', 'end', 'decemb', '2003', '.', 'eurotunnel',
"'s", 'share', ',', 'suspend', 'last', 'week', '113.5', 'penc', 'ahead',
'monday', "'s", 'announc', ',', 'resum', 'trade', 'tuesday', '.', 'sharehold',
'225', 'creditor', 'bank', 'agre', 'deal', '.', '``', 'i', "'m", 'hope', 'i',
"'m", 'take', '(', 'approv', ')', 'grant', ',', "''", 'morton', 'admit', ',',
'``', 'sharehold', 'pretti', 'angri', 'franc', '.', "''", 'ask', 'would',
'happen', 'bank', 'reject', 'deal', ',', 'morton', 'said', ',', '``', 'nobodi',
'want', 'collaps', ',', 'nobodi', 'want', 'doomsday', 'scenario', '.', "''",
'(', '$', '1=.6393', 'pound', ')']

Text 2 Tokens
----------------
['anglo-french', 'channel', 'tunnel', 'oper', 'eurotunnel', 'monday', 'announc',
'deal', 'give', 'creditor', 'bank', '45.5', 'percent', 'compani', 'return',
'wipe', 'one', 'billion', 'pound', '(', '$', '1.56', 'billion', ')', 'debt',
'.', 'the', 'long-await', 'restructur', 'bring', 'end', 'month', 'wrangl',
'eurotunnel', '225', 'bank', 'owe', 'nearli', 'nine', 'billion', 'pound', '(',
'$', '14.1', 'billion', ')', '.', 'the', 'deal', ',', 'announc', 'simultan',
'pari', 'london', ',', 'bring', 'compani', 'back', 'brink', 'insolv', 'leav',
'sharehold', 'own', '54.5', 'percent', 'compani', '.', '``', 'the',
'restructur', 'plan', 'provid', 'eurotunnel', 'medium-term', 'financi',
'stabil', 'allow', 'consolid', 'substanti', 'commerci', 'achiev', 'date',
'develop', 'oper', ',', "''", 'eurotunnel', 'co-', 'chairman', 'alastair',
'morton', 'said', '.', 'the', 'firm', 'make', 'profit', 'interest', ',', 'ad',
'.', 'although', 'sharehold', 'see', 'interest', 'dilut', ',', 'offer',
'prospect', 'brighter', 'futur', 'month', 'uncertainti', 'eurotunnel', 'wrestl',
'reduc', 'crippl', 'interest', 'payment', 'negoti', 'tunnel', "'s", 'construct',
'.', 'eurotunnel', ',', 'taken', 'around', 'half', 'cross-channel', 'market',
'european', 'ferri', 'compani', ',', 'said', 'strong', 'oper', 'perform',
'could', 'allow', 'pay', 'first', 'dividend', 'within', 'next', '10', 'year',
'.', 'french', 'co-chairman', 'patrick', 'ponsol', 'said', 'sharehold', 'would',
'patient', 'could', 'reap', 'benefit', 'compani', "'s", 'success', '.', 'he',
'call', 'debt', 'restructur', 'plan', '``', 'accept', 'compromis', "''",
```

```
'holder', 'eurotunnel', 'share', '.', 'the', 'compani', 'said', 'still',
'consider', 'work', 'done', 'finalis', 'agre', 'detail', 'plan', 'submit',
'sharehold', 'full', '225', 'bank', 'syndic', 'approv', ',', 'probabl', 'earli',
'1997', '.', 'monday', "'s", 'announc', 'follow', 'two', 'week', 'highli',
'secret', 'negoti', 'eurotunnel', 'six', 'lead', 'bank', '.', 'thi', 'extend',
'24', '``', 'instruct', 'bank', "''", 'meet', 'late', 'last', 'week', 'london',
'.', 'eurotunnel', 'said', 'debt-for-equ', 'swap', 'would', '130', 'penc', ',',
'10.40', 'franc', ',', 'per', 'share', '.', 'that', 'consider', 'level',
'around', '160', 'penc', 'wide', 'report', 'announc', 'deal', ',', 'reduc',
'outstand', 'debt', '8.7', 'billion', 'pound', '(', '$', '13.6', 'billion', ')',
'1.0', 'billion', '(', '$', '1.56', 'billion', ')', '.', 'the', 'compani',
'said', '3.7', 'billion', 'pound', '(', '$', '5.8', 'billion', ')', 'debt',
'would', 'convert', 'new', 'financi', 'instrument', 'exist', 'sharehold',
'would', 'abl', 'particip', 'issu', '.', 'if', 'choos', 'take', 'free',
'warrant', 'entitl', 'subscrib', ',', 'eurotunnel', 'said', 'sharehold', "'",
'interest', 'may', 'reduc', '39', 'percent', 'compani', 'end', 'decemb', '2003',
'.', 'eurotunnel', "'s", 'share', ',', 'suspend', 'last', 'week', '113.5',
'penc', 'ahead', 'monday', "'s", 'announc', ',', 'resum', 'trade', 'tuesday',
',', 'compani', 'said', '.']

Text 3 Tokens
----------------
['anglo-french', 'channel', 'tunnel', 'oper', 'eurotunnel', 'monday', 'announc',
'deal', 'give', 'creditor', 'bank', '45.5', 'percent', 'compani', 'return',
'wipe', 'one', 'billion', 'pound', '(', '$', '1.56', 'billion', ')', 'debt',
'mountain', '.', 'the', 'long-await', 'restructur', 'bring', 'end', 'month',
'wrangl', 'eurotunnel', '225', 'bank', 'owe', 'nearli', 'nine', 'billion',
'pound', '.', 'the', 'deal', ',', 'announc', 'simultan', 'pari', 'london', ',',
'bring', 'compani', 'back', 'brink', 'insolv', 'leav', 'sharehold', 'own',
'54.5', 'percent', 'compani', '.', '``', 'the', 'restructur', 'plan', 'provid',
'eurotunnel', 'medium', 'term', 'financi', 'stabil', 'allow', 'consolid',
'substanti', 'commerci', 'achiev', 'date', 'develop', 'oper', ',', "''",
'eurotunnel', 'co-', 'chairman', 'alastair', 'morton', 'said', '.', 'the',
'firm', 'make', 'profit', 'interest', ',', 'ad', '.', 'although', 'sharehold',
'see', 'interest', 'dilut', ',', 'offer', 'prospect', 'brighter', 'futur',
'month', 'uncertainti', 'eurotunnel', 'wrestl', 'reduc', 'crippl', 'interest',
'payment', 'negoti', 'tunnel', "'s", 'construct', '.', 'eurotunnel', ',',
'taken', 'around', 'half', 'cross-channel', 'market', 'european', 'ferri',
'compani', ',', 'said', 'strong', 'oper', 'perform', 'could', 'allow', 'pay',
'first', 'dividend', 'within', 'next', '10', 'year', '.', 'french', 'co-
chairman', 'patrick', 'ponsol', 'said', 'sharehold', 'would', 'patient',
'could', 'reap', 'benefit', 'compani', "'s", 'success', '.', 'he', 'call',
'debt', 'restructur', 'plan', '``', 'accept', 'compromis', "''", 'holder',
'eurotunnel', 'share', '.', 'the', 'compani', 'said', 'statement', 'still',
'consider', 'work', 'done', 'finalis', 'agre', 'detail', 'plan', 'submit',
'sharehold', 'full', '225', 'bank', 'syndic', 'approv', ',', 'probabl', 'earli',
'1997', '.', 'monday', "'s", 'announc', 'follow', 'two', 'week', 'highli',
'secret', 'negoti', 'eurotunnel', 'six', 'lead', 'bank', '.', 'thi', 'extend',
```

```
'24', '``', 'instruct', 'bank', "''", 'meet', 'late', 'last', 'week', 'london',
'.', 'eurotunnel', 'said', 'debt-for-equ', 'swap', 'would', '130', 'penc', ',',
'10.40', 'franc', ',', 'per', 'share', '.', 'that', 'consider', 'level',
'around', '160', 'penc', 'wide', 'report', 'run', 'deal', ',', 'reduc',
'outstand', 'debt', '8.7', 'billion', 'pound', '1.0', 'billion', '.', 'the',
'compani', 'said', '3.7', 'billion', 'pound', 'debt', 'would', 'convert', 'new',
'financi', 'instrument', 'exist', 'sharehold', 'would', 'abl', 'particip',
'issu', '.', 'if', 'choos', 'take', 'free', 'warrant', 'entitl', 'subscrib',
',', 'eurotunnel', 'said', 'sharehold', "'", 'interest', 'may', 'reduc', '39',
'percent', 'compani', 'end', 'decemb', '2003', '.', 'eurotunnel', "'s", 'share',
',', 'suspend', 'last', 'week', '113.5', 'penc', 'ahead', 'monday', "'s",
'announc', ',', 'resum', 'trade', 'tuesday', ',', 'compani', 'said', '.', '(',
'$', '1=.6393', 'pound', ')']
```

1.3 Part 3

Calculate tf-idf for each word in each document and generate document-word matrix

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

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

The tf-idf equation gives a greater weight to rare words and words that are frequent in a specific document.

The code below initializes a tf-idf vectorizer using a custom tokenizer that tokenizes the words, removes stop words, and uses the porter stemmer. The code then fits the vectorizer to the entire corpus and then transforms each individual document to a tf-idf vector.

[ ]:
```
from sklearn.feature_extraction.text import TfidfVectorizer

def tokenizer(text):
    stop_words = set(stopwords.words("english"))
    tokens = nltk.word_tokenize(text)
    stemmer = PorterStemmer()
    stemmed_tokens = [stemmer.stem(token) for token in tokens if token.lower()␣
 ↪not in stop_words]
    return stemmed_tokens
```

```
documents = [
    text_1,
    text_2,
    text_3
]

tfidf_vectorizer = TfidfVectorizer(tokenizer=tokenizer)

tfidf_vectorizer.fit(documents)

t1 = tfidf_vectorizer.transform([documents[0]])
t2 = tfidf_vectorizer.transform([documents[1]])
t3 = tfidf_vectorizer.transform([documents[2]])

feature_names = tfidf_vectorizer.get_feature_names_out()
```

The code below prints the tf-idf table.

[ ]:
```
from prettytable import PrettyTable
import numpy as np

table = PrettyTable()

table.field_names = ["Terms", "t1", "t2", "t3"]

i = len(t1.data) - 1
j = len(t2.data) - 1
k = len(t3.data) - 1

v1 = [[],[],[],[],[],[],[],[],[],[],[],[],[],[],[]]
v2 = [[],[],[],[],[],[],[],[],[],[],[],[],[],[],[]]
v3 = [[],[],[],[],[],[],[],[],[],[],[],[],[],[],[]]

count = 0

for x in range(max(t1.indices[0], t2.indices[0], t3.indices[0]), -1, -1):

  if x in t1.indices:
    d1 = t1.data[i]
    i = i-1
  else:
    d1 = 0

  if x in t2.indices:
    d2 = t2.data[j]
    j = j-1
  else:
```

```
    d2 = 0

  if x in t3.indices:
    d3 = t3.data[k]
    k = k-1
  else:
    d3 = 0

  table.add_row([feature_names[x], d1, d2, d3])

  index = int(np.floor(count/19))

  v1[index].append(d1)
  v2[index].append(d2)
  v3[index].append(d3)

  count += 1


print("Document-Word Matrix")
print(table)
```

Document-Word Matrix

Terms	t1	t2	t3
year	0.035382792235089076	0.13360547718093338	0.056160214573572474
wrestl	0.017691396117544538	0.026721095436186674	0.028080107286786237
wrangl	0	0.08016328630856003	0.08424032186035871
would	0.15922256505790083	0.13360547718093338	0.1404005364339312
work	0.0599082749131794	0.13360547718093338	0.056160214573572474
within	0.07076558447017815	0.13360547718093338	0.056160214573572474
wipe	0.0884569805877227	0.4008164315428001	0.42120160930179357
wide	0.0884569805877227	0.4008164315428001	0.42120160930179357
well	0.5838160718789698	0	0
week	0.0299541374565897	0.026721095436186674	0.028080107286786237
warrant	0.3538279223508908	0.06881655109890135	0.036158250745717255
want	0.017691396117544538	0	0
urg	0.0299541374565897	0	0
unpaid	0.0299541374565897	0	0
uncertainti	0.017691396117544538	0.026721095436186674	0.028080107286786237
two	0	0.026721095436186674	0.028080107286786237
tunnel	0.017691396117544538	0.026721095436186674	0.028080107286786237
tuesday	0.017691396117544538	0.04524274739925269	0.028080107286786237
trade	0.017691396117544538	0.026721095436186674	0.028080107286786237
told	0.0299541374565897	0	0
time	0.017691396117544538	0	0
throw	0.017691396117544538	0	0
term	0.022780893617205138	0	0.028080107286786237
ten	0.0299541374565897	0	0
taken	0.017691396117544538	0.04524274739925269	0.036158250745717255
take	0.0299541374565897	0.026721095436186674	0.028080107286786237
syndic	0	0.026721095436186674	0.056160214573572474
swap	0.017691396117544538	0.026721095436186674	0.036158250745717255
sustain	0.017691396117544538	0	0
suspend	0.017691396117544538	0.05344219087237335	0.028080107286786237

success	0	0.034408275549450675	0.028080107286786237
substanti	0	0.026721095436186674	0.028080107286786237
subscrib	0.0599082749131794	0.026721095436186674	0.028080107286786237
submit	0.017691396117544538	0.026721095436186674	0.036158250745717255
strong	0.017691396117544538	0.04524274739925269	0.08424032186035871
still	0.15922256505790083	0.026721095436186674	0.028080107286786237
statement	0.0299541374565897	0	0.036158250745717255
stabilis	0.035382792235089076	0	0
stabil	0.035382792235089076	0.034408275549450675	0.036158250745717255
spring	0.0299541374565897	0	0
six	0	0.08016328630856003	0.028080107286786237
simultan	0.0299541374565897	0.026721095436186674	0.028080107286786237
sharehold	0.035382792235089076	0.034408275549450675	0.028080107286786237
share	0.017691396117544538	0.034408275549450675	0.028080107286786237
seven	0.017691396117544538	0	0
seen	0.017691396117544538	0	0
see	0.0299541374565897	0.026721095436186674	0.056160214573572474
secur	0.017691396117544538	0	0
secret	0	0.026721095436186674	0.028080107286786237
scenario	0.0599082749131794	0	0
said	0.017691396117544538	0.026721095436186674	0.028080107286786237
run	0.0299541374565897	0	0.11232042914714495
rout	0.05307418835263361	0	0
roll	0.0299541374565897	0	0
revenu	0.035382792235089076	0	0
return	0.017691396117544538	0.026721095436186674	0.028080107286786237
resum	0.0299541374565897	0.05344219087237335	0.056160214573572474
restructur	0.0299541374565897	0.026721095436186674	0.028080107286786237
report	0.0299541374565897	0.026721095436186674	0.1404005364339312
reject	0.0299541374565897	0	0
reduc	0.017691396117544538	0.13360547718093338	0.036158250745717255
reap	0	0.026721095436186674	0.16848064372071742
provid	0.12383977282281176	0.05344219087237335	0.028080107286786237

prospect	0.0299541374565897	0.026721095436186674	0.056160214573572474
profit	0.0299541374565897	0.13360547718093338	0.028080107286786237
probabl	0.0599082749131794	0.034408275549450675	0.036158250745717255
pretti	0.0884569805877227	0	0
pound	0.0299541374565897	0.26721095436186676	0.036158250745717255
posit	0.017691396117544538	0	0
ponsol	0.017691396117544538	0.026721095436186674	0.028080107286786237
plan	0.017691396117544538	0.05344219087237335	0.028080107286786237
perform	0.0299541374565897	0.026721095436186674	0.036158250745717255
percent	0.0299541374565897	0.034408275549450675	0.028080107286786237
per	0.0299541374565897	0.034408275549450675	0.028080107286786237
penc	0.0599082749131794	0.026721095436186674	0.25272096558107615
payment	0.017691396117544538	0.026721095436186674	0.036158250745717255
pay	0.017691396117544538	0.034408275549450675	0.056160214573572474
patrick	0.035382792235089076	0.026721095436186674	0.028080107286786237
patient	0.0299541374565897	0.026721095436186674	0.028080107286786237
particip	0.0599082749131794	0.24048985892568006	0.028080107286786237
pari	0.017691396117544538	0.034408275549450675	0.056160214573572474
own	0.21229675341053444	0.05344219087237335	0.028080107286786237
owe	0	0.026721095436186674	0.028080107286786237
outstand	0	0.026721095436186674	0.028080107286786237
oper	0.0299541374565897	0.026721095436186674	0.036158250745717255
one	0.0599082749131794	0.05344219087237335	0.08424032186035871
offer	0.035382792235089076	0.026721095436186674	0.11232042914714495
note	0.017691396117544538	0	0
nobodi	0.0299541374565897	0	0
nine	0.017691396117544538	0.026721095436186674	0.028080107286786237
next	0.017691396117544538	0.026721095436186674	0.028080107286786237
news	0.0299541374565897	0	0

new	0.05307418835263361	0.034408275549450675	0.028080107286786237
negoti	0.035382792235089076	0.08016328630856003	0.028080107286786237
nearli	0.017691396117544538	0.1068843817447467	0.028080107286786237
name	0.017691396117544538	0	0
n't	0.0599082749131794	0	0
mountain	0	0	0.028080107286786237
morton	0.12383977282281176	0.026721095436186674	0.028080107286786237
month	0.10614837670526722	0.026721095436186674	0.028080107286786237
monday	0.017691396117544538	0.026721095436186674	0.056160214573572474
million	0.017691396117544538	0	0
meet	0.0299541374565897	0.026721095436186674	0.028080107286786237
medium-term	0	0.026721095436186674	0
medium	0.0299541374565897	0	0.028080107286786237
may	0.035382792235089076	0.026721095436186674	0.3088811801546486
maximum	0.017691396117544538	0	0
massiv	0.0299541374565897	0	0
market	0.017691396117544538	0.026721095436186674	0.028080107286786237
make	0.035382792235089076	0.026721095436186674	0.036158250745717255
major	0.017691396117544538	0	0
look	0.0299541374565897	0	0
long-await	0.0299541374565897	0.05344219087237335	0.028080107286786237
london	0.05307418835263361	0.026721095436186674	0.028080107286786237
like	0.0299541374565897	0	0
lifelin	0.017691396117544538	0	0
level	0.017691396117544538	0.026721095436186674	0.056160214573572474
less	0.0299541374565897	0	0
leav	0.017691396117544538	0.29393204979805343	0.028080107286786237

least	0.1769139611754454	0	0
lead	0	0.026721095436186674	0.028080107286786237
late	0	0.034408275549450675	0.036158250745717255
last	0.017691396117544538	0.026721095436186674	0.028080107286786237
keep	0.017691396117544538	0	0
issu	0.017691396117544538	0.026721095436186674	0.028080107286786237
invest	0.0299541374565897	0	0
interest	0.035382792235089076	0.05344219087237335	0.028080107286786237
instrument	0.0299541374565897	0.026721095436186674	0.036158250745717255
instruct	0	0.026721095436186674	0.028080107286786237
insolv	0	0.034408275549450675	0.028080107286786237
ingeni	0.017691396117544538	0	0
hope	0.017691396117544538	0	0
holder	0	0.026721095436186674	0.028080107286786237
hold	0.0299541374565897	0	0
highli	0.0299541374565897	0.026721095436186674	0.028080107286786237
help	0.0299541374565897	0	0
happen	0.0299541374565897	0	0
half	0.035382792235089076	0.026721095436186674	0.036158250745717255
group	0.017691396117544538	0	0
grant	0.017691396117544538	0	0
go	0.035382792235089076	0	0
give	0.0299541374565897	0.034408275549450675	0.036158250745717255
get	0.0299541374565897	0	0
game	0.017691396117544538	0	0
futur	0.0299541374565897	0.026721095436186674	0.036158250745717255
full	0	0.026721095436186674	0.028080107286786237
french	0.0299541374565897	0.026721095436186674	0.11232042914714495
free	0.0299541374565897	0.026721095436186674	0.028080107286786237

franc	0.017691396117544538	0.034408275549450675	0.056160214573572474
forward	0.0299541374565897	0	0
formula	0.0299541374565897	0	0
follow	0	0.034408275549450675	0.036158250745717255
flow	0.017691396117544538	0	0
fix	0.0299541374565897	0	0
first	0.0299541374565897	0.034408275549450675	0.036158250745717255
firm	0.0299541374565897	0.026721095436186674	0.028080107286786237
find	0.017691396117544538	0	0
financi	0.12383977282281176	0.1068843817447467	0.028080107286786237
financ	0.0599082749131794	0	0
finalis	0.05307418835263361	0.026721095436186674	0.056160214573572474
ferri	0.0299541374565897	0.05344219087237335	0.028080107286786237
extend	0	0.034408275549450675	0.028080107286786237
exist	0.017691396117544538	0.034408275549450675	0.028080107286786237
eurotunnel	0.0299541374565897	0.026721095436186674	0.028080107286786237
european	0.017691396117544538	0.026721095436186674	0.036158250745717255
equiti	0.0299541374565897	0	0
entitl	0.017691396117544538	0.05344219087237335	0.028080107286786237
end	0.0299541374565897	0.026721095436186674	0.08424032186035871
eight	0.0299541374565897	0	0
earli	0.017691396117544538	0.026721095436186674	0.056160214573572474
dwindl	0.017691396117544538	0	0
doomsday	0.0299541374565897	0	0
done	0.0299541374565897	0.026721095436186674	0.028080107286786237
dividend	0.017691396117544538	0.026721095436186674	0.04754375448244289
dilut	0.017691396117544538	0.04524274739925269	0.028080107286786237
difficult	0.0299541374565897	0	0

develop	0.0299541374565897	0.026721095436186674	0.056160214573572474
detail	0.017691396117544538	0.08016328630856003	0.028080107286786237
despit	0.022780893617205138	0	0
depreci	0.017691396117544538	0	0
decemb	0.08986241236976909	0.05344219087237335	0.028080107286786237
debt-for-equ	0.035382792235089076	0.026721095436186674	0.028080107286786237
debt	0.017691396117544538	0.026721095436186674	0.028080107286786237
deal	0.07076558447017815	0.05344219087237335	0.028080107286786237
date	0	0.026721095436186674	0.08424032186035871
current	0.0299541374565897	0	0
cross-channel	0.0299541374565897	0.026721095436186674	0.036158250745717255
crippl	0.035382792235089076	0.026721095436186674	0.036158250745717255
creditor	0.017691396117544538	0.026721095436186674	0.028080107286786237
could	0.017691396117544538	0.026721095436186674	0.028080107286786237
cost	0.0599082749131794	0	0
convert	0.017691396117544538	0.08016328630856003	0.028080107286786237
construct	0.035382792235089076	0.034408275549450675	0.028080107286786237
constitut	0.0599082749131794	0	0
consolid	0.08986241236976909	0.034408275549450675	0.028080107286786237
consider	0.017691396117544538	0.026721095436186674	0.028080107286786237
confer	0.017691396117544538	0	0
compromis	0	0.026721095436186674	0.028080107286786237
complex	0.05307418835263361	0	0
compani	0.017691396117544538	0.026721095436186674	0.08424032186035871
commerci	0.035382792235089076	0.026721095436186674	0.028080107286786237
come	0.017691396117544538	0	0
collaps	0.017691396117544538	0	0

co-chairman	0.017691396117544538	0.026721095436186674	0.08424032186035871
co-	0	0.026721095436186674	0.028080107286786237
choos	0.05307418835263361	0.026721095436186674	0.08424032186035871
channel	0.035382792235089076	0.08016328630856003	0.028080107286786237
chairman	0	0.026721095436186674	0.1404005364339312
cash	0.05307418835263361	0	0
cap	0.017691396117544538	0	0
call	0	0.08016328630856003	0.028080107286786237
ca	0.05307418835263361	0	0
busiest	0.035382792235089076	0	0
brink	0.035382792235089076	0.026721095436186674	0.028080107286786237
bring	0.017691396117544538	0.08016328630856003	0.028080107286786237
brighter	0.0299541374565897	0.026721095436186674	0.028080107286786237
bond	0.1415311689403563	0	0
billion	0.0299541374565897	0.1068843817447467	0.036158250745717255
bill	0.017691396117544538	0	0
benefit	0	0.026721095436186674	0.08424032186035871
becom	0.017691396117544538	0	0
bankruptci	0.017691396117544538	0	0
bank	0.017691396117544538	0.026721095436186674	0.028080107286786237
back	0.035382792235089076	0.026721095436186674	0.08424032186035871
avoid	0.0299541374565897	0	0
avail	0.05307418835263361	0	0
ask	0.017691396117544538	0	0
arrang	0.017691396117544538	0	0
around	0.017691396117544538	0.026721095436186674	0.028080107286786237
approv	0.0299541374565897	0.034408275549450675	0.028080107286786237
annual	0.0299541374565897	0	0

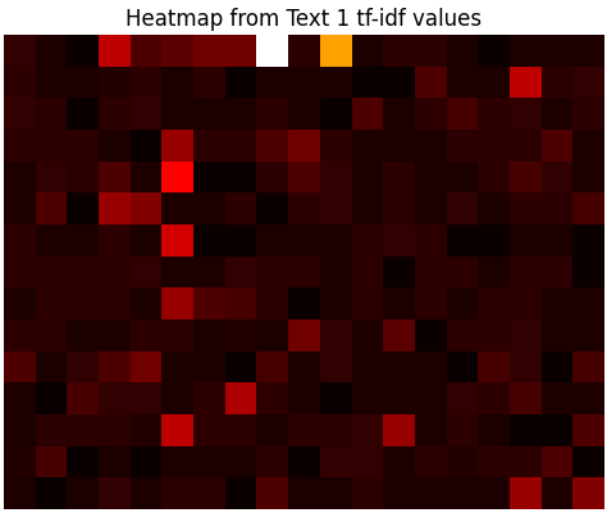
announc	0.0299541374565897	0.08016328630856003	0.036158250745717255
angri	0.022780893617205138	0	0
anglo-french	0.15922256505790083	0.026721095436186674	0.2246408582942899
analyst	0.0299541374565897	0	0
although	0.0299541374565897	0.08016328630856003	0.036158250745717255
alreadi	0.017691396117544538	0	0
allow	0.0299541374565897	0.026721095436186674	0.028080107286786237
alastair	0.0299541374565897	0.026721095436186674	0.08424032186035871
ahead	0.035382792235089076	0.2137687634894934	0.16848064372071742
agre	0.12383977282281176	0.034408275549450675	0.028080107286786237
afford	0.017691396117544538	0	0
admit	0.0299541374565897	0	0
ad	0.017691396117544538	0.026721095436186674	0.036158250745717255
achiev	0	0.08016328630856003	0.028080107286786237
accept	0	0.16032657261712005	0.036158250745717255
abl	0.0599082749131794	0.026721095436186674	0.028080107286786237
abil	0.022780893617205138	0	0
``	0.05307418835263361	0.034408275549450675	0.028080107286786237
8.7	0	0.026721095436186674	0.028080107286786237
54.5	0.017691396117544538	0.026721095436186674	0.028080107286786237
5.8	0	0.026721095436186674	0
45.5	0.017691396117544538	0.026721095436186674	0.036158250745717255
400	0.017691396117544538	0	0
39	0.017691396117544538	0.026721095436186674	0.036158250745717255
3.7	0.0299541374565897	0.034408275549450675	0.028080107286786237
24	0	0.034408275549450675	0.028080107286786237
225	0.035382792235089076	0.026721095436186674	0.036158250745717255
2004	0.035382792235089076	0	0

2003.	0.017691396117544538	0.026721095436186674	0.028080107286786237
200	0.0299541374565897	0	0
1=.6393	0.022780893617205138	0	0.028080107286786237
1997.	0.0299541374565897	0.034408275549450675	0.036158250745717255
160	0.0299541374565897	0.026721095436186674	0.028080107286786237
150	0.0599082749131794	0	0
14.1	0	0.026721095436186674	0
130	0.017691396117544538	0.026721095436186674	0.028080107286786237
13.6	0	0.026721095436186674	0
113.5	0.017691396117544538	0.05344219087237335	0.056160214573572474
10.40	0.035382792235089076	0.034408275549450675	0.036158250745717255
10	0.017691396117544538	0.026721095436186674	0.028080107286786237
1.85	0.0299541374565897	0	0
1.6	0.0299541374565897	0	0
1.56	0	0.026721095436186674	0.028080107286786237
1.0	0.0599082749131794	0.08016328630856003	0.08424032186035871
.	0.017691396117544538	0.026721095436186674	0.028080107286786237
--	0.017691396117544538	0	0
,	0.0299541374565897	0.026721095436186674	0.028080107286786237
)	0.017691396117544538	0.026721095436186674	0.028080107286786237
(	0.017691396117544538	0.026721095436186674	0.028080107286786237
's	0.017691396117544538	0.1068843817447467	0.11232042914714495
'm	0.017691396117544538	0	0
''	0.12383977282281176	0.034408275549450675	0.036158250745717255
'	0.017691396117544538	0.026721095436186674	0.028080107286786237
$	0.10614837670526722	0.026721095436186674	0.028080107286786237

The code below generates a heatmap using the tf-idf vector of text 1.

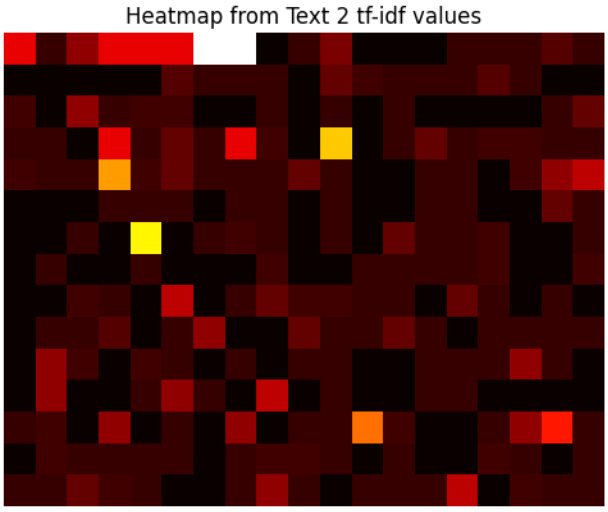
[128]:
```
import matplotlib.pyplot as plt

plt.imshow(v1, cmap='hot', interpolation='nearest')
plt.title('Heatmap from Text 1 tf-idf values')
plt.axis('off')
plt.show()
```



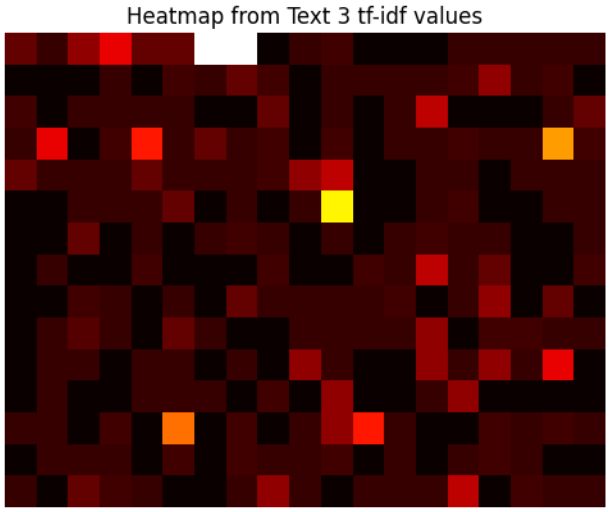
The code below generates a heatmap using the tf-idf vector of text 2.

[129]:
```
plt.imshow(v2, cmap='hot', interpolation='nearest')
plt.axis('off')
plt.title('Heatmap from Text 2 tf-idf values')
plt.show()
```



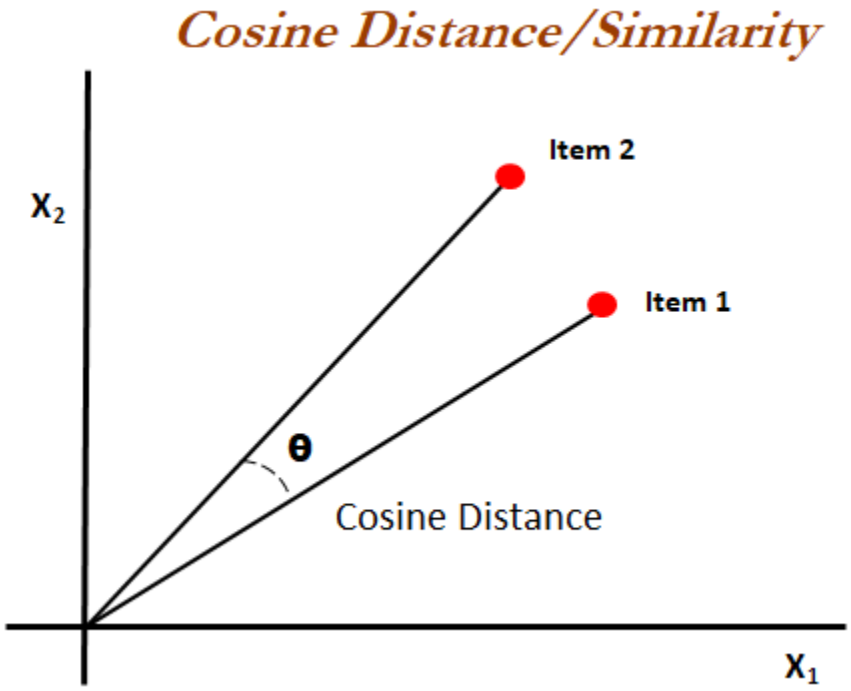
The code below generates a heatmap using the tf-idf vector of text 3.

[130]:
```
plt.imshow(v3, cmap='hot', interpolation='nearest')
plt.axis('off')
plt.title('Heatmap from Text 3 tf-idf values')
plt.show()
```



1.4 Part 4

Calculate pairwise cosine similarity for the documents



The cosine similarity of two document's tf-idf vectors represents how similar they are.

The code below calculates the cosine similarity of each document and prints it.

[134]:
```
from sklearn.metrics.pairwise import cosine_similarity

print("Cosine similarity between t1 and t2:")
print("------------------------------------")
print(cosine_similarity(t1, t2)[0][0])
print()

print("Cosine similarity between t1 and t3:")
print("------------------------------------")
print(cosine_similarity(t1, t3)[0][0])
print()

print("Cosine similarity between t2 and t3:")
print("------------------------------------")
print(cosine_similarity(t2, t3)[0][0])
```

```
Cosine similarity between t1 and t2:
------------------------------------
0.8187743943854953
```

```
Cosine similarity between t1 and t3:
------------------------------------
0.8391995483557352

Cosine similarity between t2 and t3:
------------------------------------
0.9752383249870626
```

1.5 Discussions and Conclusions

Overall, documents 2 and 3 are deemed most similar through the cosine similarity of their tf-idf vectors, which aligns with the intuition gained by reading all three documents. Documents 2 and 3 are nearly the same, expect for a few different words, which aligns with the 0.97 similarity score. Document 1 contains more differences from 2 and 3, which aligns with the ~0.82-0.84 similarity scores respectively. Heatmaps 2 and 3 also appear more similar than 1 and 2 or 1 and 3, which supports the cosine similarity scores obtained.

1.6 References

- https://scikit-learn.org/stable/index.html
- https://www.nltk.org/
- https://python.plainenglish.io/introduction-to-nltk-library-in-python-6fa729b54ad
- https://medium.com/@ajay_khanna/tokenization-techniques-in-natural-language-processing-67bb22088c75
- https://ted-mei.medium.com/demystify-tf-idf-in-indexing-and-ranking-5c3ae88c3fa0
- https://www.oreilly.com/library/view/statistics-for-machine/9781788295758/eb9cd609-e44a-40a2-9c3a-f16fc4f5289a.xhtml