PEMPROSESAN DOKUMEN *PORTABLE DOCUMENT FORMAT (PDF)* PANGKALAN KORPUS DEWAN BAHASA DAN PUSTAKA

Nurul Athirah binti Ahmad Sabri

Dr. Sabrina Binti Tiun

Fakulti Teknologi dan Sains Maklumat, Universiti Kebangsaan Malaysia

ABSTRAK

Pangkalan Data Korpus berfungsi sebagai sebuah pangkalan data untuk menempatkan data korpus Bahasa Melayu dalam bentuk digital. Justeru, pihak Dewan Bahasa dan Pustaka telah mengambil inisiatif untuk menyediakan pelbagai kemudahan infrastruktur kepada pengguna apabila menyedari kemajuan teknologi maklumat yang kian berkembang pesat. Pihak Dewan Bahasa dan Pustaka memerlukan penambahbaikan terhadap sistem yang sedia ada bagi memenuhi keperluan pengguna lantaran pertambahan hasil kepustakaan korpus dalam pangkalan data. Penambahbaikan dalam sistem diperlukan untuk meningkatkan efisensi kepada pengguna, manakala untuk pihak Dewan Bahasa dan Pustaka pula, sistem yang dapat menjimatkan masa dan pantas diperlukan untuk memproses bahan- bahan kepustakaan agar lebih efisien. Dalam proses penambahbaikan sudah tentu mempunyai beberapa masalah yang dihadapi. Masalah yang diberi perhatian adalah proses penandaan dokumen secara manual oleh Dewan Bahasa Pustaka terhadap dokumen jenis "Portable Document Format (PDF)". Hal ini akan menyebabkan kekangan masa untuk memproses bahan kepustakaan yang banyak. Membangunkan algoritma untuk sistem proses penandaan secara automatik merupakan penyelesaian yang dikemukakan untuk mengatasi masalah ini. Menggunakan Metodologi "Waterfall Methodology" kerana membudayakan perancangan yang teliti dan strategik untuk mengurangkan permasalahan yang bakal terjadi semasa membangunkan projek kajian. Kesimpulannya, dengan adanya projek ini, pihak Dewan Bahasa Dan Pustaka dapat memberi fasiliti yang lebih baik kepada pengguna.

1 PENGENALAN

Bahan "Portable Documen Format" merupakan bahan utama yang akan digunakan untuk kajian ini. Bahan-bahan yang disimpan di dalam korpus terdiri daripada bentuk tulisan atau lisan. Setiap jenis wacana diklasifikasikan mengikut subkorpus yang berasingan. Pangkalan

Data Korpus mempunyai saiz (sehingga 25 November 2008) adalah lebih kurang 135 juta perkataan yang terkandung dalam sepuluh kategori subkorpus. Oleh kerana pertambahan bahan yang banyak, pihak Dewan Bahasa dan Pustaka memerlukan satu algoritma untuk memudahkan mereka menguruskan serta memproses bahan dengan cepat, terutama bahan jenis "Portable Document Format(PDF)" untuk disimpan di dalam korpus. Algoritma yang akan dibangunkan akan membuat penandaan bahan "Portable Document Format (PDF)" secara automatik sekali gus akan mengurangkan penggunaan masa dan tenaga untuk memproses bahan "Portable Document Format (PDF)". Hal ini kerana pihak Dewan Bahasa dan Pustaka masih membuat penandaan secara manual di mana proses memakan masa yang lama. Kepentingan penandaan terhadap bahan "Portable Document Format" untuk mengekstrak informasi mengikut klasifikasi misalnya seperti muka surat, perenggan dan ayat. Pengekstrakan berlaku terhadap dua jenis teks iaitu teks jenis format pantun dan teks normal. Informasi tesebut amat penting semasa proses penyimpanan data ke dalam storan korpus.

2 PENYATAAN MASALAH

Pihak Dewan Bahasa dan Pustaka masih membuat penandaan terhadap bahan "Portable Document Format (PDF)" secara manual untuk disimpan di dalam Pangkalan Data Korpus melalui tenaga pekerja yang telah dilatih. Hal ini akan menyebabkan bahan yang akan disimpan di dalam korpus semakin banyak dek kerana kekangan masa dan tenaga untuk membuat penandaan. Hasil penandaan mengeluarkan maklumat bilangan perenggan, perkataan dan ayat. Oleh itu, pihak Dewan Bahasa dan Pustaka perlu mengambil pelbagai inisiatif untuk mempercepatkan proses penandaan dalam menyediakan servis yang bagus kepada pelayar internet.

3 OBJEKTIF KAJIAN

Projek ini bertujuan membangunankan pemprosesan Portable Document Format (PDF) bagi memudahkan pengguna untuk mengekstrak maklumat. Projek ini diharapkan dapat membantu pengguna memproses dokumen dengan lebih cepat dan tepat. Secara umum objektif kajian adalah menghasilkan alatan penandaan secara automatik untuk mengekstrak informasi daripada dokumen “Portable Document Format (PDF)” ke bentuk format berstruktur seperti .txt. Selain itu, objektif kajian adalah membina algoritma yang mudah difahami untuk memproses dokumen “Portable Dokumen Format (PDF)” untuk pengguna. Akhir sekali, penyimpanan dan pengurusan output objektif pertama ke dalam format .csv agar lebih sistematik dan kemas agar nanti mudah diintegrasikan dengan sistem korpus DBP.

4 METOD KAJIAN

Penggunaan model pembangunan yang sesuai amat penting untuk memastikan perjalanan projekberjalan dengan lancar dan memperoleh hasil seperti yangdirancang. Pemprosesan dokumen pangkalan korpus DBP akan dibangunkan menggunakan pendekatan kajian melalui model Kitar Hayat Pembangunan Sistem (System Development Life Cycle) iaitu Model Air Terjun. Hal ini kerana model ini mempunyai dokumentasi dan tahap perancangan yang bergerak ke arah matlamat dan tujuan yang sama secara konsisten. Keupayaan model ini mampu membimbing sepasukan ke arah matlamat yang sama kerana mudah difahami dalam mengatur tugas. Selain itu, metod ini menggunakan konsep menyelesaikan langkah demi langkah dibuat untuk memastikan setiap fasa dihabiskan sebelum ke fasa seterusnya **.**

4.1 Fasa Perancangan

Fasa Perancangan adalah satu fasa yang penting dalam membangunkan pemprosesan dokumen ini kerana menampakkan dengan jelas tentang penyelidikan awal terhadap projek kajian ini. Hal ini kerana, pelbagai masalah yang timbul dalam menjalankan projek yang akan dikenal pasti semasa dalam fasa perancangan. Pada fasa perancangan, permasalahan, objektif serta skop dapat dikenal pasti. Selain itu, carian serta pengumpulan sorotan sastera yang berkaitan dengan kajian dijalankan melalui kajian - kajian lepas dan buku. Maklumat dikumpul, dan digunakan dalam fasa analisis.

4.2 Fasa Analisis

Fasa Analisis dijalankan untuk menganalisis latar belakang pembangunan projek kajian serta mengenal pasti kekurangan atau kelemahan terhadap sistem yang sedia ada. Sistem – sistem yang sedia ada diselidik untuk dijadikan bahan rujukan dalam mengumpul dan menganalisis maklumat tersebut dengan efektif. Sistem – sistem telah dianalisis untuk membangunkan sistem yang lebih baik serta memenuhi keperluan dan kehendak pengguna. Di samping itu, analisis tentang perisian dan perkakasan juga dijalankan untuk memastikan kesusuaian dengan projek kajian.

4.3 Fasa Reka Bentuk

Fasa rekabentuk merupakan proses bagi pembangunan projek kajian yang telah dirancang berdasarkan hasil kajian daripada fasa analisis. Fasa ini mengenalpasti bagaimana algoritma penandaan secara automatik ini berfungsi dari aspek perisian dan reka bentuk algoritma. Dalam fasa ini, proses merekabentuk algoritma dilakukan. Pemilihan terhadap bahasa pengaturcaraan serta perisian untuk menjalankan sistem ini dikenal pasti. Pilihan platform yang digunakan adalah Jupyter Notebook dan menggunakan bahasa Python.

4.4 Fasa Implemetasi

Pada Fasa implementasi, penglibatan bahasa pengaturcaraan diimplimentasikan ke dalam pemprosesan bagi *Python* seperti dirancang. . Selain itu, proses pengekstrakan menggunakan *library PyPDF*. Selain itu, terdapat proses segmentasi dijalankan dalam kajian menggunakan kaedah *split()* untuk mengklasifikasikan dokumen kepada perenggan, ayat dan perkataan.

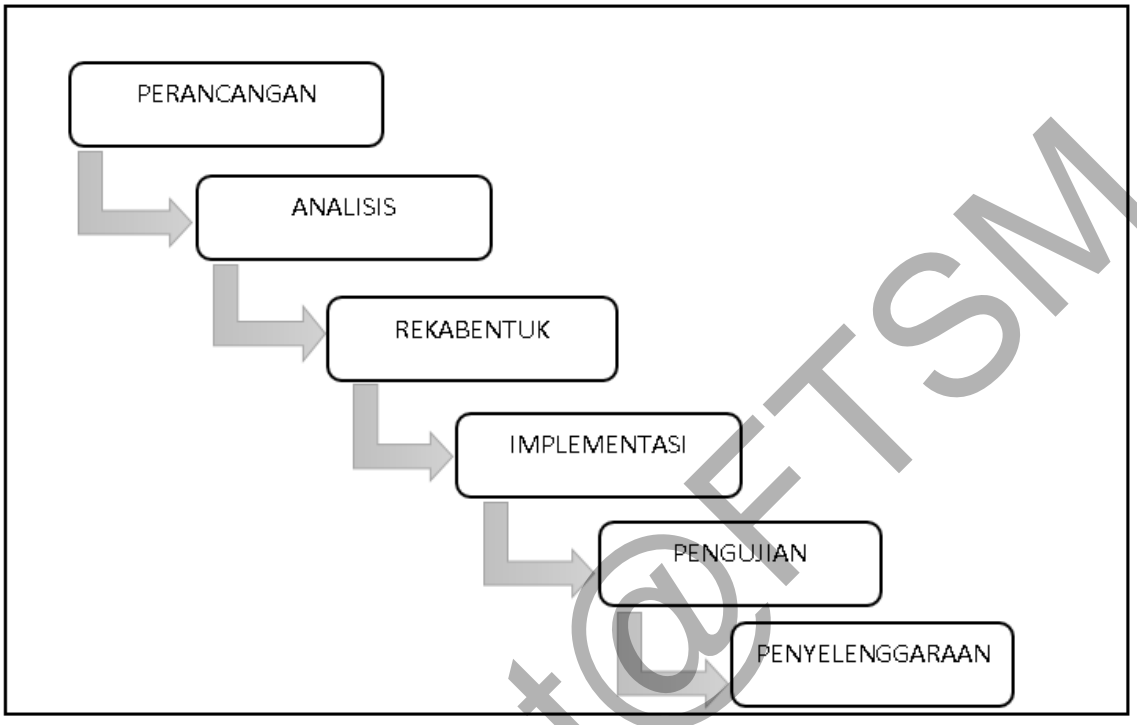
4.5 Fasa Pengujian

Fasa Pengujian akan dijalankan selepas fasa implementasi berfungsi bertujuan menguji keberkesanan dan efisen projek terhadap bahan yang diterima daripada pihak pengguna supaya tidak mengalami sebarang ralat atau kesulitan yang akan memberi kesan terhadap kelancaran sistem ini. Pemprosesan ini diuji untuk mengkelaskan dokumen PDF kepada perenggan, ayat dan perkataan. Fasa Penyelenggaraan merupakan fasa terakhir bagi model Air Terjun ini. Pemantauan dilakukan secara berkala bagi mengekalkan prestasi sistem pemprosesan supaya tiada sebarang masalah berlaku di luar kawalan. Untuk projek ini, fasa pemantauan tidak dilakukan kerana algoritma untuk sistem akan sentiasa berubah mengikut syarat teks. pengaturcaraan serta perisian untuk menjalankan sistem ini dikenal pasti. Pilihan platform yang digunakan adalah Jupyter Notebook dan menggunakan bahasa Python.

4.6 Fasa Penyelenggaraan

Fasa Penyelenggaraan merupakan fasa terakhir bagi model Air Terjun ini. Pemantauan dilakukan secara berkala bagi mengekalkan prestasi sistem pemprosesan supaya tiada sebarang masalah berlaku di luar kawalan. Untuk projek ini, fasa pemantauan tidak dilakukan kerana algoritma untuk sistem akan sentiasa berubah mengikut syarat teks. pengaturcaraan serta

perisian untuk menjalankan sistem ini dikenal pasti. Pilihan platform yang digunakan adalah Jupyter Notebook dan menggunakan bahasa Python.



Rajah 1 Model Pembangunan Pemprosesan Dokumen Portable Dokumen Format (PDF) Pangkalan Korpus Dewan Bahasa Dan Pustaka

5 HASIL KAJIAN

Bahagian ini akan membincang hasil daripada proses – proses yang terlibat dalam pemprosesan PDF. Dalam projek ini, perisisan Anaconda digunakan untuk membangunkan algoritma. Jupyter Notebook digunakan sebagai platform untuk membina algoritma yang akan membuat penandaan secara automati terhadap dokumen yang dipilih. Seterusnya, pengujian dilakukan terhadap pemprosesan untuk menguji kebolehan algoritma untuk membezaan perenggan, ayat dan perkataan. Pengujian dijalankan untuk memastikan hasil daripada pembangunan adalah selaras dengan objektif yang telah dirancang. Proses – proses yang terlibat akan melibatkan

dua jenis format dalam kajian ini. Format yang terlibat adalah format pantun dan format teks normal. Terdapat lima proses yang akan terlibat dalam pemprosesan dokumen PDF.

Rajah 2 menunjukkan hasil algoritma yang berfungsi untuk mengektrak dokumen dari dokumen jenis. PDF ke .TXT akan digunakan untuk proses selanjutnya. Format ini diekstrak supaya memudahkan pemprosesan teks dan lebih teratur menggunakan *library PyPDF*. Skop kajian juga dicapai kerana skop kajian ini adalah memproses bahan jenis PDF.

```
!pip install PyPDF2
import PyPDF2 as pdf
file = open('ikankekek.pdf', 'rb')
file
pdf_reader = pdf.PdfFileReader(file)
pdf_reader

page1 =pdf_reader.getPage(0)
page1.extractText()
```

```
Requirement already satisfied: PyPDF2 in c:\programdata\anaconda3\lib\site-packages (1.26.0)

'1BAB 1PengenalanIkan kekek mak ilui-ilui,  \nIkan gelama mak ilai-ilai, \nNanti adik mak ilui-ilui, \nPulang sama mak ila
i-ilai.  \n Tenang-tenang air di laut,  \nSampan golek mudik ke tanjung,  \nHati terkenang mulut tersebut,  \nBudi yang ba
ik rasa dijunjung.  \n Ikan kekek mak ilui-ilui,  \nIkan gelama mak ilai-ilai,  \nNanti adik mak ilui-ilui,  \nPulang sama
mak ilai-ilai.Dari mana punai melayang,Dari paya turun ke padi,Dari mana datangnya sayang,dari mata turun ke hati.Lagu fiIk
an Kekekfl merupakan salah satu lagu yang biasanya \ndinyanyikan oleh masyarakat Melayu ketika hendak menidurkan \nanak kec
il mereka. Lagu tersebut banyak menerapkan nilai-nilai \nikan kekek_25092018.indd   19/26/18   9:30 AM'
```

Rajah 2 Proses Pengekstrakan Teks PDF

Rekabentuk algoritma menunjukkan hasil pemprosesan dijalankan terhadap format teks pantun yang telah disegmentasikan kepada perenggan (A), ayat (B) dan perkataan(C) ditunjukkan dalam Rajah 3. Proses segmentasi tersebut menggunkan kaedah split() berdasarkan klu yang jelas untuk setiap dokumen.

(A)

```
ayat = '1BAB 1PengenalanIkan kekek mak ilui-ilui,  \nIkan gelama mak ilai-ilai, \nNanti adik mak ilui-ilui, \nPulang sama
def sentences_pattern(text) :
    string_split = ayat.split("\n")
    print(string_split)
sentences_pattern(ayat)
```

```
ayat = '1BAB 1PengenalanIkan kekek mak ilui-ilui,  \nIkan gelama mak ilai-ilai, \nNanti adik mak ilui-ilui, \nPulang sama m
def word_pattern(text) :
    string_split = ayat.split(" ")
    print(string_split)
word_pattern(ayat)
```

```
['1BAB', '1PengenalanIkan', 'kekek', 'mak', 'ilui-ilui,', '', '\nIkan', 'gelama', 'mak', 'ilai-ilai,', '\nNanti', 'adik',
'mak', 'ilui-ilui,', '\nPulang', 'sama', 'mak', 'ilai-ilai.', '', '\n', 'Tenang-tenang', 'air', 'di', 'laut,', '', '\nSamp
an', 'golek', 'mudik', 'ke', 'tanjung,', '', '\nHati', 'terkenang', 'mulut', 'tersebut,', '', '\nBudi', 'yang', 'baik', 'r
asa', 'dijunjung.', '', '\n', 'Ikan', 'kekek', 'mak', 'ilui-ilui,', '', '\nIkan', 'gelama', 'mak', 'ilai-ilai,', '', '\nNa
nti', 'adik', 'mak', 'ilui-ilui,', '', '\nPulang', 'sama', 'mak', 'ilai-ilai.Dari', 'mana', 'punai', 'melayang,Dari', 'pay
a', 'turun', 'ke', 'padi,Dari', 'mana', 'datangnya', 'sayang,dari', 'mata', 'turun', 'ke', 'hati.']
```

(B)

(C)

Rajah 3 Proses Segmentasi Terhadap Teks Pantun

```
pantun = '1BAB 1PengenalanIkan kekek mak ilui-ilui,  \nIkan gelama

def pantun_pattern (pantuntext) :
    string_split = pantun.split(".")
    print(string_split)

pantun_pattern(pantun)
```

```
['1BAB 1PengenalanIkan kekek mak ilui-ilui,  \nIkan gelama mak ilai
-ilai, \nNanti adik mak ilui-ilui, \nPulang sama mak ilai-ilai', '
\n Tenang-tenang air di laut,  \nSampan golek mudik ke tanjung,  \n
Hati terkenang mulut tersebut,  \nBudi yang baik rasa dijunjung', '
\n Ikan kekek mak ilui-ilui,  \nIkan gelama mak ilai-ilai,  \nNanti
adik mak ilui-ilui,  \nPulang sama mak ilai-ilai', 'Dari mana punai
melayang,Dari paya turun ke padi,Dari mana datangnya sayang,dari ma
ta turun ke hati']
```

Bagi memudahkan pengguna melihat hasil pemprosesan dengan lebih kemas, peggunaan *dataframe* diperluaskan kepada setiap proses. Rajah 4 dibawah menunjukkan hasil pemprosesan selepas proses segmentasi dan disusun rapi.

```
df.style.set_properties(subset=['Perenggan'], **{'width': '600px'})
```

	Perenggan
0	1BAB 1PengenalanIkan kekek mak ilui-ilui, Ikan gelama mak ilai-ilai, Nanti adik mak ilui-ilui, Pulang sama mak ilai-ilai
1	Tenang-tenang air di laut, Sampan golek mudik ke tanjung, Hati terkenang mulut tersebut, Budi yang baik rasa dijunjung
2	Ikan kekek mak ilui-ilui, Ikan gelama mak ilai-ilai, Nanti adik mak ilui-ilui, Pulang sama mak ilai-ilai
3	Dari mana punai melayang,Dari paya turun ke padi,Dari mana datangnya sayang,dari mata turun ke hati

```
DATA = {'Perenggan':['Dimensi\n \nBahasa\n \ndalam\n \nAgama\n \n \nBAHASA dan agama ialah dua perkara yang saling melengka
df = pd.DataFrame(DATA)
df['Perenggan'] = df['Perenggan'].str.replace(r'\W'," ")
df.to_csv('textnormalperenggan.csv')
print(df)
```

```
                                          Perenggan
0  Dimensi   Bahasa   dalam   Agama     BAHASA da...
1  Menurut penulis  bahasa juga merupakan teras p...
2  Dimensi bahasa Arab menjadi lebih luas apabila...
3  Namun  tidak dapat dinafikan hari ini ahli bah...
4  Seterusnya penulis menyebut bahawa hubungan ba...
```

Rajah 4 Menunjukkan Hasil Pemprosesan Mengunakan Dataframe

Rekabentuk algoritma menunjukkan hasil pemprosesan dijalan terhadap format teks normal yang telah disegmentasikan kepada perenggan (A), ayat (B) dan perkataan(C) ditunjukkan dalam Rajah 5. Proses segmentasi tersebut menggunkan kaedah split() berdasarkan klu yang jelas untuk setiap dokumen. Teks normal boleh ditanda menggunakan klu yang kerap digunakan seperti noktah untuk mengasingkan ayat. Selepas itu, dokumen yang telah siap diproses akan disimpan dalam .csv untuk memudahkan pengguna mengemaskini dokumen.

(A)

```
import pandas as pd
DATA = {'Perkataan':['Dimensi\n', '\nBahasa\n', '\ndalam\n', '\nAgama\n', '\n', '\nBAHASA', 'dan', 'agama', 'ialah', 'dua',
df = pd.DataFrame(DATA)
df['Perkataan'] = df['Perkataan'].str.replace(r'\W'," ")
df.to_csv('perkataannormal.csv')
print(df)
```

```
     Perkataan
0      Dimensi
1       Bahasa
2        dalam
3        Agama
4
..         ...
376    tamadun
377   Manakala
378    agamawan
379        pula
380

[381 rows x 1 columns]
```

(B)

(C)

Rajah 5 Proses Segmentasi Terhadap Teks Pantun

```
import pandas as pd
DATA = {'Ayat':['Dimensi\n \nBahasa\n \ndalam\n \nAgama\n \n \nBAHASA dan agama ialah dua perkara yang saling melengkapi an
df = pd.DataFrame(DATA)
df['Ayat'] = df['Ayat'].str.replace(r'\W'," ")
df.to_csv('ayatnormal.csv')

print(df)
```

```
                                                Ayat
0   Dimensi   Bahasa   dalam   Agama     BAHASA da...
1    Setiap  bahasa mempunyai sejarah dalam penyeb...
2    Semua kitab suci   agama ditulis dalam  bahas...
3    Oleh itu  bahasa memainkan peranan yang penti...
4     Menurut penulis  bahasa juga merupakan tera...
5    Manusia dalam tamadun tersebut  mempunyai pro...
6    Con tohnya tamadun Islam yang dibina  oleh Ra...
7    Secara rasminya tamadun Islam lahir melalui r...
8      Dimensi bahasa Arab menjadi lebih luas apab...
9    Maka  memahami bahasa Arab dengan baik menjad...
10    Justeru itu  perlu difahami bahawa bahasa Ar...
11     Namun  tidak dapat dinafikan hari ini ahli ...
12   Walaupun usaha penterjemahan bahasa suatu  ya...
13   Seorang yang beragama Islam tidak cukup memah...
14     Seterusnya penulis menyebut bahawa hubungan...
15                         Manakala  agamawan pula
```

6 KESIMPULAN

Kajian ini dijalankan untuk memproses teks bagi sebuah dokumen format Portable Document Format (PDF). Objektif kajian ini adalah untuk membina alatan penandaan secara automatik berasaskan algoritma yang dibina. Dengan adanya pemprosesan ini, pengurusan output yang sistematik dapat dihasilkan menerusi penggunaan algoritma yang dinyatakan. Oleh itu, dengan pembangunan pemprosesan ini, dokumen dapat diproses dengan lancar.

Dalam kajian ini, beberapa teknik pemprosesan data telah digunakan. Perisian Python digunakan dalam kajian ini untuk membina algoritma yang menpunyai pustaka atau *library* yang meluas seperti *PyPDF* untuk mengekstrak *PDF*. Hasil daripada pembangunan pemprosesan ini, algoritma yang dibina dapat diimplementasikan ke dalam pemprosesan dokumen seperti mengklasifikasikan perenggan, ayat, perkataan secara automatik menggunkan perisian anaconda. Sistem ini dibangunkan menggunakan bahasa pengaturcaraan *Python* yang disokong oleh perisian *anaconda* yang memudahkan untuk memproses teks.

Secara keseluruhan, pemprosesan ini dapat mencapai objektif serta keperluan pengguna seperti dirancang. Matlamat untuk pembinaan berdasarkan sumber terbuka berjaya dilaksanakan. Bagi membangunkan sebuah sistem yang bagus, perancangan yang teliti serta disiplin yang konsisten dan pengguna metodologi yang sesuai perlu dititi beratkan. Diharapkan pemprosesan ini dapat diterima pengguna serta memberi manfaat kepada pihak-pihak yang terlibat.

7 RUJUKAN

Kiril Simov, Alexander Simov, Milen Kouylekov, Krasimira Ivanova, Ilko Grigorov, Hristo Geneva. (2003). *Development* of *Corpora within the CLaRK System: The BulTreeBank Project Experience.* In: Proc. of the Demo Sessions of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL'03), Budapest, Hungary.

Shahbaz, Muhammad, et al. (2014) “Sentiment Miner: A Prototype for Sentiment Analysis of Unstructured Data and Text.” 2014 IEEE 27th Canadian Conference on Electrical and Computer Engineering (CCECE), doi:10.1109/ccece.2014.6901087.

Amazon Simple Storage Service and Amazon Glacier Storage. (2019). AWS Certified SolutionsArchitect Study Guide, 47–66. doi: 10.1002/9781119560395.ch3
AWS Global Infrastructure.Machine Learning in the AWS Cloud, 151–160. doi: 10.1002/9781119556749.ch7.

Richards, Mark. (2015) “Software Architecture Patterns.” O'Reilly | Safari, O'Reilly Media, “Text Mining: Converting Between Tidy & Non-Tidy Formats.” Text Mining: Converting Between Tidy & Non-Tidy Formats · AFIT Data Science Lab R Programming Guide,

Feldman R. and Hirsh H. (1997) Finding Associations in Collections of Text. In Michalski R.S., Bratko I. and Kubat M. (edts) Machine Learning, Data Mining and Knowledge Discovery: Methods and Application (John Wiley and sons Ltd).

Y. Wilks (1997). Information extraction as a core language technology. In M-T. Pazienza, editor, Information Extraction. Springer, Berlin.

U. Nahm and R. Mooney (2002). Text mining with information extraction. In Proceedings of the AAAI 2002 Spring Symposium on Mining Answers from Texts and Knowledge Bases.

R. Gaizauskas (2003). An information extraction perspective on text mining: Tasks, technologies and prototype applications.
http://www.itri.bton.ac.uk/projects/euromap/TextMiningEvent/Rob_Gaizauskas.pdf.