PENGECAMAN ENTITI NAMA PINTAR DARI TEKS INGGERIS

Nurul Camelia Binti Murad

Prof. Madya Dr Mohd Zakree Bin Ahmad

Fakulti Teknologi dan Sains Maklumat, Universiti Kebangsaan Malaysia

ABSTRAK

Pengecaman Entiti Nama (PEN) adalah satu proses dimana untuk mengenal pasti entiti nama dalam sesebuah dokumen dan mengklasifikasikannya mengikut kategori iaitu individu, lokasi, organisasi, tarikh atau masa. Bahasa Inggeris dijadikan bahasa utama yang akan digunakan dalam projek ini. Teks dalam berbentuk korpus yang akan dimasukkan oleh pengguna akan digunakan untuk mengasingkan dan mengklasifikasikan entiti nama mengikut kategori. Sistem ini akan menggunakan sepenuhnya bahasa pengaturcaraan Python.

1 PENGENALAN

Pada masa ini, terdapat banyak teks daripada maklumat dalam keadaan yang tidak berstruktur seperti teks dari surat khabar, e-mel, blog serta kicauan dari media sosial. Pengekstrakan maklumat merupakan salah satu kaedah yang boleh digunakan untuk mewakili sesuatu kandungan teks dari dokumen yang tidak berstruktur.

Pemprosesan Bahasa Tabii (PBT) merupakan salah satu bidang penting dalam Sains Komputer (Alfred et al. 2014). PBT akan menganalisis teks berdasarkan kedua-dua set teori dan teknologi (Liddy 2001). PBT menjadikan maklumat tidak berstruktur kepada maklumat yang lebih bermakna. Pengekstrakan Maklumat (PM) merupakan proses mengestrak maklumat daripada artikel yang tidak berstruktur untuk menjadikannya hasil maklumat yang berguna (Alfred et al. 2014). Salah satu cabang dalam PM ialah Pengecaman Entiti Nama (PEN).Proses PEN ini sangat berguna dalam mengenal pasti dan mengecam entiti seperti individu, organisasi atau lokasi.

2 PENYATAAN MASALAH

Maklumat serta dokumen yang tidak berstruktur terutamanya daripada media sosial sukar difahami oleh komputer. Komputer juga tidak dapat mengenal pasti kata nama khas dalam korpus bagi sesetengah dokumen. Proses untuk mengekstrak info atau kata kunci daripada dokumen tidak berstruktur tersebut secara manual memerlukan masa yang lama terutama dokumen yang mengandungi teks dalam skala yang besar. Masalah ini menjadi lebih kompleks apabila penulis menggunakan singkatan untuk mewakili sesuatu kata nama. Persoalannya adalah bagaimana meningkatkan keberkesanan alat pengecam nama entiti bahasa Inggeris?

3 OBJEKTIF KAJIAN

Objektif projek ini adalah seperti berikut :-

i. Membina korpus bahasa Inggeris baharu dari media sosial dan Web.

ii. Membangunkan model pengecaman entiti nama.

iii. Membangunkan prototaip aplikasi pengecaman nama entiti berasaskan Web.

4 METOD KAJIAN

Metodologi yang digunakan dalam proses membangunkan sistem ini adalah dengan menggunakan kitaran pembangunan perisian jenis Agile. Kelebihan menggunakan kitaran tersebut kerana boleh membahagikan projek ini kepada beberapa peringkat dan dapat membuat penambahbaikan secara berterusan dalam setiap peringkat. Selain itu, perisian kerja lebih penting daripada dokumen secara menyeluruh. Maklum balas daripada pengguna juga sangat penting dalam rekabentuk ini untuk memastikan sebarang kesilapan dalam projek dapat dibetulkan pada setiap fasa. Pengguna boleh melihat dan memberi idea apa yang pengguna mahukan berdasarkan setiap fasa tanpa perlu menunggu pada akhir projek.

1. Fasa Analisis Keperluan (*Requirement specification*)

Fasa ini adalah untuk mengumpul maklumat berkaitan projek ini untuk membangunkan Pengecaman Entiti Nama Pintar Dari Teks Inggeris. Fasa ini juga akan membincangkan kajian yang akan dilakukan terhadap pengguna untuk membangunkan perisian yang mengikut kehendak pengguna.

2. Fasa Reka Bentuk (*Design*)

Fasa reka bentuk akan mula untuk mengenal pasti apakah perisian atau platform yang akan digunakan bagi membina sistem dalam projek ini. Selain itu, fasa ini mereka bentuk jenis antara muka yang mesra pengguna secara lakaran.

3. Fasa Pembangunan (*Integration*)

Fasa pembangunan akan menggabungkan semua idea, keperluan, perisian yang diperoleh daripada fasa-fasa sebelum ini untuk membangunkan sistem. Pengaturcaraan juga merupakan salah satu keperluan yang paling penting dalam fasa ini. Antara muka juga akan

diimplementasikan dalam sistem ini. Sistem akan dibangunkan sepenuhnya dan seterusnya diuji dalam fasa pengujian.

4. Fasa Pengujian (*Testing and debugging*)

Fasa pengujian bertujuan untuk menguji sistem yang telah dibangunkan dalam fasa sebelum ini. Semua fungsi di dalam sistem tersebut akan diuji. Segala data, masalah yang berlaku akan dikumpul untuk fasa penyelenggaraan.

5. Fasa Penyelenggaraan (*Maintenance*)

Dalam fasa yang terakhir ini, semua data dan masalah yang dikumpul akan dirujuk dan digunakan untuk memperbaiki sistem kepada yang lebih baik.

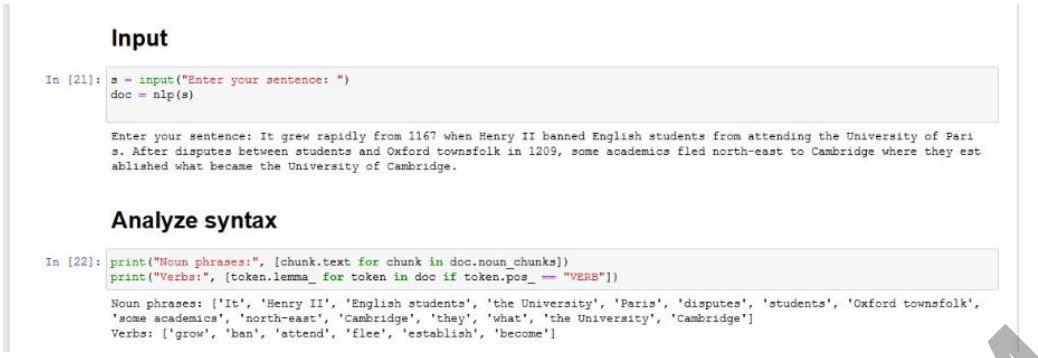
5 HASIL KAJIAN

Bahagian ini membincang hasil daripada proses pembangunan sistem Pengecaman Entiti Nama Pintar Dari Teks Inggeris. Perisian yang digunakan untuk membangunkan system ini adalah dengan menggunakan *Jupyter Notebook.* Selain itu, terdapat pemasang pakej yang digunakan ketika membangunkan projek ini iaitu *pip*. Sistem ini menggunakan bahasa *Python* sebagai bahasa utama.

Terdapat dua bahagian untuk input iaitu khusus kepada ayat pendek dan dokumen.

5.1 Ayat Pendek

Rajah 5.1 di bawah menujukkan kotak panel untuk diisi ayat pendek oleh pengguna. Kotak panel kedua di bawah akan menganalisis sintaks untuk setiap perkataan dan dibahagi kepada kata nama dan juga kata kerja.



Input

```
In [21]: s = input("Enter your sentence: ")
         doc = nlp(s)

         Enter your sentence: It grew rapidly from 1167 when Henry II banned English students from attending the University of Pari
         s. After disputes between students and Oxford townsfolk in 1209, some academics fled north-east to Cambridge where they est
         ablished what became the University of Cambridge.
```

Analyze syntax

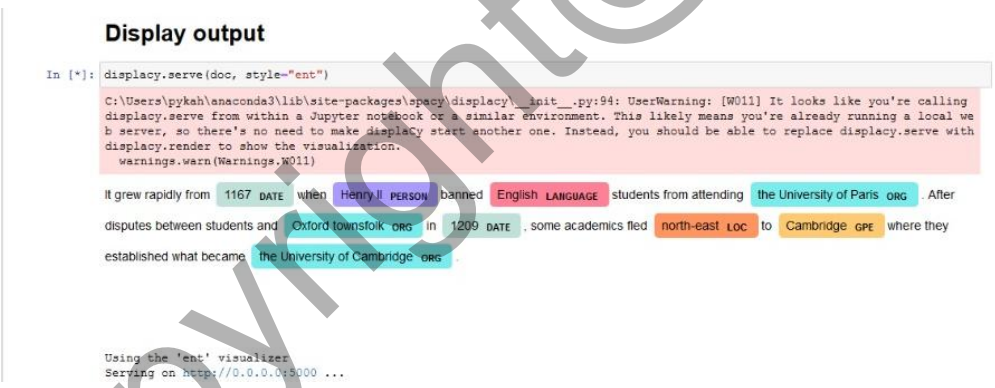
```
In [22]: print("Noun phrases:", [chunk.text for chunk in doc.noun_chunks])
         print("Verbs:", [token.lemma_ for token in doc if token.pos_ == "VERB"])

         Noun phrases: ['It', 'Henry II', 'English students', 'the University', 'Paris', 'disputes', 'students', 'Oxford townsfolk',
         'some academics', 'north-east', 'Cambridge', 'they', 'what', 'the University', 'Cambridge']
         Verbs: ['grow', 'ban', 'attend', 'flee', 'establish', 'become']
```

(Rajah 5.2 Input dan analisis sintaks)

Seterusnya, rajah 5.3 menunjukkan sistem menganalisis dan menyenarai pendek frasa atau perkataan yang dikenal pasti sebagai entiti nama. Di sebelah perkataan pula menunjukkan jenis entiti nama tersebut tergolong.

(Rajah 5.2 Pencarian frasa dan entiti nama)

Rajah 5.3 menunjukkan hasil akhir bagi input ayat oleh pengguna. Perkataan yang telah diwarnakan merupakan entiti nama setelah diproses.



(Rajah 5.4 Hasil akhir)

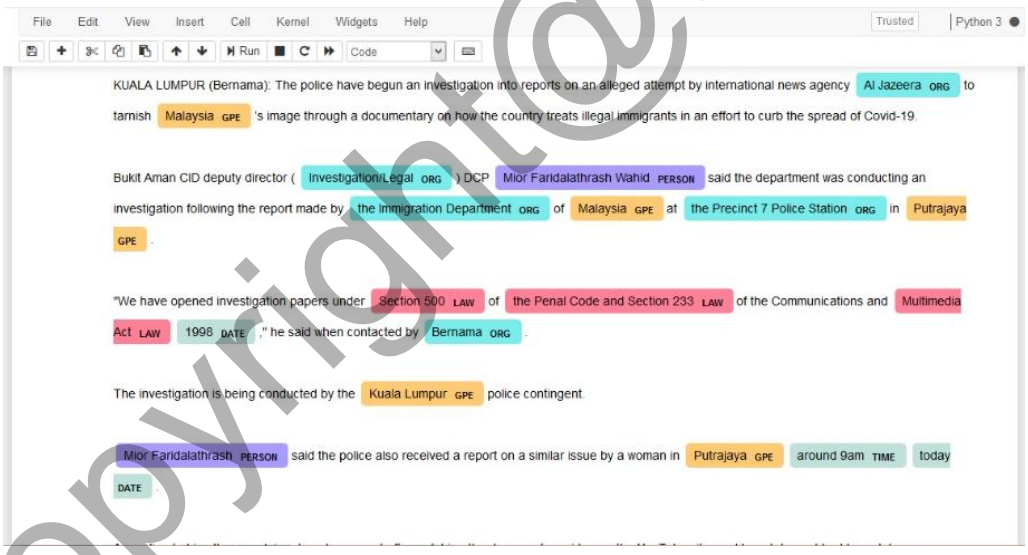
5.2 Dokumen

Pengguna juga boleh menganalisis teks dokumen dalam berbentuk *Word(docx.)* untuk diproses dalam sistem ini. Rajah 5.5 menunjukkan kod untuk membaca teks dokumen oleh pengguna.

```
In [3]: import docx2txt
        doc = docx2txt.process("C:/Users/pykah/Desktop/document/jazeera.docx")
        print(doc)

Police begin investigation into Al Jazeera report on illegal immigrants in Malaysia.


Monday, 06 Jul 2020

10:32 PM MYT


KUALA LUMPUR (Bernama): The police have begun an investigation into reports on an alleged attempt by international news age
ncy Al Jazeera to tarnish Malaysia's image through a documentary on how the country treats illegal immigrants in an effort
to curb the spread of Covid-19.

Bukit Aman CID deputy director (Investigation/Legal) DCP Mior Faridalathrash Wahid said the department was conducting an in
vestigation following the report made by the Immigration Department of Malaysia at the Precinct 7 Police Station in Putraja
ya.

"We have opened investigation papers under Section 500 of the Penal Code and Section 233 of the Communications and Multimed
ia Act 1998," he said when contacted by Bernama.

The investigation is being conducted by the Kuala Lumpur police contingent.

Mior Faridalathrash said the police also received a report on a similar issue by a woman in Putrajaya around 9am today.

According to him, the complainant made a report after watching the documentary video on the YouTube site, and her statement
had been taken.
```

(Rajah 5.4 Input teks dokumen oleh pengguna)

Langkah seterusnya sama seperti dalam rajah 5.1, 5.2 dan 5.2. Rajah 5.5 di bawah menunjukkan contoh hasil akhir entiti nama bagi input teks dokumen.



(Rajah 5.6 Hasil akhir teks dokumen)

6 KESIMPULAN

Kesimpulannya, sistem Pengecaman Entiti Nama Pintar Dari Teks Inggeris ini telah dibangunkan mengikut perancangan yang telah dirancang berpandukan metodologi kajian. Sebahagian objektif dan penyelesaian masalah telah berjaya diselesaikan walaupun masih mempunyai sedikit ralat yang perlu diperbaiki serta objektif yang tidak berjaya untuk dicapai. Diharapkan kewujudan sistem ini akan dapat memberi manfaat kepada penggunanya.

7 RUJUKAN

Alfred, R., Leong, L.C., On, C.K. & Anthony, P. 2014. Malay Named Entity Recognition Based on Rule-Based Approach. *International Journal of Machine Learning and Computing4*(3): 300–306.

Dey, A., Syam, B. & Professor, P. 2013. Named Entity Recognition using Gazetteer Method and N-gram Technique for an Inflectional Language: A Hybrid Approach. *International Journal of Computer Applications* Vol. 84

Django Architecture - 3 Major Components of MVC Pattern - DataFlair. (t.th.). https://data-flair.training/blogs/django-architecture/ [9 October 2019].

Grishman, R. (t.th.). Message Understanding Conference-6: A Brief History.

Jahangir, F., Anwar, W., Ijaz Bajwa, U. & Wang, X. 2012. N-gram and Gazetteer List Based Named Entity Recognition for Urdu: A Scarce Resourced Language.

Liddy, E.D. 2001. Natural Language Processing Natural Language Processing Natural Language Processing 1. https://surface.syr.edu/istpub [5 December 2019].

Mustafa, S.H. & Al-Radaideh, Q.A. 2004. Using N-grams for Arabic text searching. *Journal of the American Society for Information Science and Technology55*(11): 1002–1007.

Nadeau, D. & Sekine, S. (t.th.). A survey of named entity recognition and classification. http://projects.ldc.upenn.edu/gale/ [5 December 2019].

Saad, S. & Mansor, M.K. 2018. Pendekatan Teknik Pengecaman Entiti Nama Bagi Capaian Berita Jenayah Bahasa Melayu (Named Entity Recognition Approach for Malay Crime

News Retrieval). *GEMA Online® Journal of Language Studies18*(4): 216–235. http://ejournal.ukm.my/gema/article/view/28999/8687 [5 December 2019].