

Ezra Pasha Ramadhansyah - 2006597872

1. Perbandingan

Menjalankan notebook:

Vocab Size 500

Nomor Kalimat	Tokenisasi BPE	Tokenisasi WordPiece	Gold Standard
1	'Pem', 'erintah', 'ko', 'ta', 'D', 'el', 'h', 'i', 'meng', 'er', 'ah', 'kan', 'm', 'on', 'y', 'et', 'untuk', 'meng', 'u', 'si', 'r', 'm', 'on', 'y', 'et', '-', 'm', 'on', 'y', 'et', 'la', 'in', 'yang', 'ber', 'ba', 'dan', 'lebih', 'ke', 'c', 'il', 'dari', 'arena', 'P', 'es', 'ta', 'O', 'lah', 'r', 'aga', 'Per', 'sem', 'ak', 'm', 'ur', 'an', '.'	'P', '##e', '##m', '##e', '##r', '##i', '##n', '##t', '##a', '##h', 'k', '##o', '##t', '##a', 'D', '##e', '##i', '##h', '##i', 'm', '##e', '##n', '##g', '##e', '##r', '##a', '##h', '##k', '##a', '##n', 'm', '##o', '##n', '##y', '##e', '##t', 'u', '##n', '##t', '##u', '##k', 'm', '##e', '##n', '##g', '##u', '##s', '##i', '##r', 'm', '##o', '##n', '##y', '##e', '##t', '-', 'm', '##o', '##n', '##y', '##e', '##t', 'l', '##a', '##i', '##n', 'y', '##a', '##n', '##g', 'b', '##e', '##r', '##b', '##a', '##d', '##a', '##n', 'l', '##e', '##b', '##i', '##h', 'k', '##e', '##c', '##i', '##i', 'd', '##a', '##r', '##i', 'a', '##r', '##e', '##n', '##a', 'P', '##e', '##s', '##t', '##a', 'O', '##i', '##a', '##h', '##r', '##a', '##g', '##a', 'P', '##e', '##r', '##s', '##e', '##m', '##a', '##k', '##m', '##u', '##r', '##a', '##n', '.'	'Pemerintah', 'kota', 'Delhi', 'mengarahkan', 'monyet', 'untuk', 'mengusir', 'monyet-monyet', 'lain', 'yang', 'berbadan', 'lebih', 'kecil', 'dari', 'arena', 'Pesta', 'Olahraga', 'Persemakmuran', '.'
2	'Pen', 'an', 'am', 'an', 'modal', 'as', 'ing', '(', 'P', 'M', 'A', ')', 'di', 'M',	'P', '##e', '##m', '##e', '##r', '##i', '##n', '##t', '##a', '##h', 'k', '##o',	'Penanaman', 'modal', 'asing', '(', 'PMA', ')', 'di',

	'al', 'ay', 'sia', 'tahun', '2006', 'mencapai', 'li', 'ma', 'k', 'ali', 'lebih', 'besar', 'dibanding', 'kan', 'Indonesia', ',', 'h', 'al', 'ini', 'men', 'un', 'j', 'uk', 'kan', 'pemb', 'ang', 'unan', 'ekonomi', 'M', 'al', 'ay', 'sia', 'ja', 'uh', 'lebih', 'men', 'ari', 'k', 'dibanding', 'kan', 'Indonesia', 'ba', 'gi', 'in', 'ves', 'tor', 'as', 'ing', '.'	'##t', '##a', 'D', '##e', '##l', '##h', '##i', 'm', '##e', '##n', '##g', '##e', '##r', '##a', '##h', '##k', '##a', '##n', 'm', '##o', '##n', '##y', '##e', '##t', 'u', '##n', '##t', '##u', '##k', 'm', '##e', '##n', '##g', '##u', '##s', '##i', '##r', 'm', '##o', '##n', '##y', '##e', '##t', '-', 'm', '##o', '##n', '##y', '##e', '##t', 'l', '##a', '##i', '##n', 'y', '##a', '##n', '##g', 'b', '##e', '##r', '##b', '##a', '##d', '##a', '##n', 'l', '##e', '##b', '##i', '##h', 'k', '##e', '##c', '##i', '##l', 'd', '##a', '##r', '##i', 'a', '##r', '##e', '##n', '##a', 'P', '##e', '##s', '##t', '##a', 'O', '##l', '##a', '##h', '##r', '##a', '##g', '##a', 'P', '##e', '##r', '##s', '##e', '##m', '##a', '##k', '##m', '##u', '##r', '##a', '##n', '.'	'Malaysia', 'tahun', '2006', 'mencapai', 'lima', 'kali', 'lebih', 'besar', 'dibandingkan', 'Indonesia', ',', 'hal', 'ini', 'menunjukkan', 'pembangunan', 'ekonomi', 'Malaysia', 'jauh', 'lebih', 'menarik', 'dibandingkan', 'Indonesia', 'bagi', 'investor', 'asing', '.'
--	--	--	---

Akurasi tokenisasi Kalimat 1

BPE = 5 / 19 = 0.263

WordPiece = 1/19 = 0.053

Akurasi tokenisasi Kalimat 2

BPE = 16 / 33 = 0.485

WordPiece = 6/33 = 0.182

Vocab Size 1000

Nomor Kalimat	Tokenisasi BPE	Tokenisasi WordPiece	Gold Standard
1	'Pemerintah', 'ko', 'ta', 'D', 'el', 'hi', 'meng', 'erah', 'kan', 'm', 'ony',	'P', '##e', '##m', '##e', '##r', '##i', '##n', '##t', '##a', '##h', 'k', '##o',	'Pemerintah', 'kota', 'Delhi', 'mengerahkan',

	<p>'et', 'untuk', 'meng', 'u', 'si', 'r', 'm', 'ony', 'et', '-', 'm', 'ony', 'et', 'lain', 'yang', 'ber', 'ba', 'dan', 'lebih', 'ke', 'c', 'il', 'dari', 'arena', 'P', 'es', 'ta', 'O', 'lah', 'r', 'aga', 'Per', 'sem', 'ak', 'm', 'uran', '.'</p>	<p>'##t', '##a', 'D', '##e', '##l', '##h', '##i', 'm', '##e', '##n', '##g', '##e', '##r', '##a', '##h', '##k', '##a', '##n', 'm', '##o', '##n', '##y', '##e', '##t', 'u', '##n', '##t', '##u', '##k', 'm', '##e', '##n', '##g', '##u', '##s', '##i', '##r', 'm', '##o', '##n', '##y', '##e', '##t', '-', 'm', '##o', '##n', '##y', '##e', '##t', 'l', '##a', '##i', '##n', 'y', '##a', '##n', '##g', 'b', '##e', '##r', '##b', '##a', '##d', '##a', '##n', 'l', '##e', '##b', '##i', '##h', 'k', '##e', '##c', '##i', '##l', 'd', '##a', '##r', '##i', 'a', '##r', '##e', '##n', '##a', 'P', '##e', '##s', '##t', '##a', 'O', '##l', '##a', '##h', '##r', '##a', '##g', '##a', 'P', '##e', '##r', '##s', '##e', '##m', '##a', '##k', '##m', '##u', '##r', '##a', '##n', '.'</p>	<p>'monyet', 'untuk', 'mengusir', 'monyet-monyet', 'lain', 'yang', 'berbadan', 'lebih', 'kecil', 'dari', 'arena', 'Pesta', 'Olahraga', 'Persemakmuran', '.'</p>
2	<p>'Pen', 'an', 'aman', 'modal', 'asing', '(', 'P', 'M', 'A', ')', 'di', 'M', 'al', 'ay', 'sia', 'tahun', '2006', 'mencapai', 'lima', 'k', 'ali', 'lebih', 'besar', 'dibandingkan', 'Indonesia', ',', 'h', 'al', 'ini', 'menunjukkan', 'pembangunan', 'ekonomi', 'M', 'al', 'ay', 'sia', 'ja', 'uh', 'lebih', 'men', 'ari', 'k', 'dibandingkan', 'Indonesia', 'bagi', 'inves', 'tor', 'asing', '.'</p>	<p>'P', '##e', '##n', '##a', '##n', '##a', '##m', '##a', '##n', 'm', '##o', '##d', '##a', '##l', 'a', '##s', '##i', '##n', '##g', '(', 'PMA', ')', 'd', '##i', 'M', '##a', '##l', '##a', '##y', '##s', '##i', '##a', 't', '##a', '##h', '##u', '##n', '2006', 'm', '##e', '##n', '##c', '##a', '##p', '##a', '##i', 'l', '##i', '##m', '##a', 'k', '##a', '##l', '##i', 'l', '##e', '##b', '##i', '##h', 'b', '##e', '##s', '##a', '##r', 'd', '##i',</p>	<p>'Penanaman', 'modal', 'asing', '(', 'PMA', ')', 'di', 'Malaysia', 'tahun', '2006', 'mencapai', 'lima', 'kali', 'lebih', 'besar', 'dibandingkan', 'Indonesia', ',', 'hal', 'ini', 'menunjukkan', 'pembangunan', 'ekonomi', 'Malaysia', 'jauh', 'lebih', 'menarik', 'dibandingkan',</p>

		'##b', '##a', '##n', '##d', '##i', '##n', '##g', '##k', '##a', '##n', 'l', '##n', '##d', '##o', '##n', '##e', '##s', '##i', '##a', ',', 'h', '##a', '##l', 'l', '##n', '##i', 'm', '##e', '##n', '##u', '##n', '##j', '##u', '##k', '##k', '##a', '##n', 'p', '##e', '##m', '##b', '##a', '##n', '##g', '##u', '##n', '##a', '##n', 'e', '##k', '##o', '##n', '##o', '##m', '##i', 'M', '##a', '##l', '##a', '##y', '##s', '##i', '##a', 'j', '##a', '##u', '##h', 'l', '##e', '##b', '##i', '##h', 'm', '##e', '##n', '##a', '##r', '##i', '##k', 'd', '##i', '##b', '##a', '##n', '##d', '##i', '##n', '##g', '##k', '##a', '##n', 'l', '##n', '##d', '##o', '##n', '##e', '##s', '##i', '##a', 'b', '##a', '##g', '##i', 'i', '##n', '##v', '##e', '##s', '##l', '##o', '##r', 'a', '##s', '##i', '##n', '##g', 'l'	'Indonesia', 'bagi', 'investor', 'asing', '.'
--	--	---	--

Akurasi tokenisasi Kalimat 1

BPE = $8 / 19 = 0.421$

WordPiece = $1/19 = 0.053$

Akurasi tokenisasi Kalimat 2

BPE = $24 / 33 = 0.727$

WordPiece = $6/33 = 0.182$

Vocab Size 5000

Nomor Kalimat	Tokenisasi BPE	Tokenisasi WordPiece	Gold Standard
---------------	----------------	----------------------	---------------

1	<p>'Pemerintah', 'kota', 'Delhi', 'meng', 'erah', 'kan', 'monyet', 'untuk', 'mengusir', 'monyet', '-', 'monyet', 'lain', 'yang', 'berbadan', 'lebih', 'kecil', 'dari', 'arena', 'P', 'es', 'ta', 'O', 'lah', 'raga', 'Per', 'sem', 'ak', 'm', 'uran', ''</p>	<p>'Pemerint', '##a', '##h', 'kot', '##a', 'Delhi', 'menger', '##a', '##h', '##k', '##a', '##n', 'monyet', 'untuk', 'mengu', '##s', '##i', '##r', 'monyet', '-', 'monyet', 'l', '##a', '##i', '##n', 'y', '##a', '##ng', 'berb', '##a', '##d', '##a', '##n', 'lebih', 'kecil', 'd', '##a', '##r', '##i', 'ar', '##e', '##n', '##a', 'Pe', '##s', '##t', '##a', 'Ol', '##a', '##h', '##r', '##a', '##g', '##a', 'Perse', '##m', '##a', '##k', '##m', '##u', '##r', '##a', '##n', ''</p>	<p>'Pemerintah', 'kota', 'Delhi', 'mengerahkan', 'monyet', 'untuk', 'mengusir', 'monyet-monyet', 'lain', 'yang', 'berbadan', 'lebih', 'kecil', 'dari', 'arena', 'Pesta', 'Olahraga', 'Persemakmuran', ''</p>
2	<p>'Pen', 'anaman', 'modal', 'asing', '(', 'PMA', ')', 'di', 'Malaysia', 'tahun', '2006', 'mencapai', 'lima', 'kali', 'lebih', 'besar', 'dibandingkan', 'Indonesia', ',', 'hal', 'ini', 'menunjukkan', 'pembangunan', 'ekonomi', 'Malaysia', 'jauh', 'lebih', 'menarik', 'dibandingkan', 'Indonesia', 'bagi', 'investor', 'asing', ''</p>	<p>'Pen', '##a', '##n', '##a', '##m', '##a', '##n', 'mod', '##a', '##l', 'asing', '(', 'PMA', ')', 'd', '##i', 'M', '##a', '##l', '##a', '##y', '##s', '##i', '##a', 't', '##a', '##h', '##u', '##n', '2006', 'menc', '##a', '##p', '##a', '##i', 'l', '##i', '##m', '##a', 'k', '##a', '##l', '##i', 'lebih', 'bes', '##a', '##r', 'd', '##i', '##b', '##a', '##n', '##d', '##i', '##ng', '##k', '##a', '##n', 'Indonesi', '##a', ',', 'h', '##a', '##l', 'ini', 'menunjukk', '##a', '##n', 'pemb', '##a', '##ng', '##u', '##n', '##a', '##n', 'ekonomi', 'M', '##a', '##l', '##a', '##y', '##s', '##i', '##a', 'j', '##a', '##u', '##h', 'lebih', 'men', '##a', '##r', '##i', '##k', 'd', '##i', '##b', '##a', '##n', '##d', '##i', '##ng', '##k', '##a', '##n', 'Indonesi', '##a', 'b', '##a', '##g', '##i',</p>	<p>'Penanaman', 'modal', 'asing', '(', 'PMA', ')', 'di', 'Malaysia', 'tahun', '2006', 'mencapai', 'lima', 'kali', 'lebih', 'besar', 'dibandingkan', 'Indonesia', ',', 'hal', 'ini', 'menunjukkan', 'pembangunan', 'ekonomi', 'Malaysia', 'jauh', 'lebih', 'menarik', 'dibandingkan', 'Indonesia', 'bagi', 'investor', 'asing', ''</p>

		'investor', 'asing', '.'	
--	--	--------------------------	--

Akurasi tokenisasi Kalimat 1

BPE = 14 / 19 = 0.737

WordPiece = 6/19 = 0.316

Akurasi tokenisasi Kalimat 2

BPE = 32 / 33 = 0.970

WordPiece = 13/33 = 0.394

Vocab Size 10000

Nomor Kalimat	Tokenisasi BPE	Tokenisasi WordPiece	Gold Standard
1	'Pemerintah', 'kota', 'Delhi', 'meng', 'erah', 'kan', 'monyet', 'untuk', 'mengusir', 'monyet', '-', 'monyet', 'lain', 'yang', 'berbadan', 'lebih', 'kecil', 'dari', 'arena', 'P', 'es', 'ta', 'O', 'lah', 'raga', 'Per', 'sem', 'ak', 'm', 'uran', '.'	'Pemerintah', 'kot', '##a', 'Delhi', 'menger', '##ahk', '##a', '##n', 'monyet', 'untuk', 'mengusir', 'monyet', '-', 'monyet', 'lain', 'yang', 'berbadan', 'lebih', 'kecil', 'd', '##a', '##ri', 'ar', '##e', '##n', '##a', 'Pe', '##s', '##t', '##a', 'Ol', '##a', '##h', '##r', '##a', '##ga', 'Perse', '##m', '##a', '##k', '##mu', '##r', '##a', '##n', '.'	'Pemerintah', 'kota', 'Delhi', 'mengerahkan', 'monyet', 'untuk', 'mengusir', 'monyet-monyet', 'lain', 'yang', 'berbadan', 'lebih', 'kecil', 'dari', 'arena', 'Pesta', 'Olahraga', 'Persemakmuran', '.'
2	'Penanaman', 'modal', 'asing', '(', 'PMA', ')', 'di', 'Malaysia', 'tahun', '2006', 'mencapai', 'lima', 'kali', 'lebih', 'besar', 'dibandingkan', 'Indonesia', ',', 'hal', 'ini', 'menunjukkan', 'pembangunan', 'ekonomi', 'Malaysia', 'jauh', 'lebih', 'menarik', 'dibandingkan', 'Indonesia', 'bagi', 'investor', 'asing', '.'	'Pen', '##a', '##n', '##a', '##m', '##a', '##n', 'modal', 'asing', '(', 'PMA', ')', 'di', 'Malaysia', 'tahun', '2006', 'mencapai', 'lima', 'kali', 'lebih', 'besar', 'dibandingkan', 'Indonesia', ',', 'hal', 'ini', 'menunjukkan', 'pembangunan', 'ekonomi', 'Malaysia', 'jauh', 'lebih', 'menarik', 'dibandingkan', 'Indonesia', 'bagi', 'investor', 'asing', '.' 'lebih', 'men', '##a', '##r', '##i', '##k', 'd', '##i', '##b', '##a', '##n',	'Penanaman', 'modal', 'asing', '(', 'PMA', ')', 'di', 'Malaysia', 'tahun', '2006', 'mencapai', 'lima', 'kali', 'lebih', 'besar', 'dibandingkan', 'Indonesia', ',', 'hal', 'ini', 'menunjukkan', 'pembangunan', 'ekonomi', 'Malaysia', 'jauh', 'lebih', 'menarik', 'dibandingkan',

		'##d', '##i', '##ng', '##k', '##a', '##n', 'Indonesi', '##a', 'b', '##a', '##g', '##i', 'investor', 'asing', '.'	'Indonesia', 'bagi', 'investor', 'asing', '.'
--	--	--	--

Akurasi tokenisasi Kalimat 1

BPE = $14 / 19 = 0.737$

WordPiece = $11/19 = 0.579$

Akurasi tokenisasi Kalimat 2

BPE = $33 / 33 = 1$

WordPiece = $32 / 33 = 0.970$

Analisis :

Kalimat 1 v Kalimat 2

Terlepas dari cara tokenisasi, kalimat 2 memiliki akurasi yang lebih tinggi dibandingkan kalimat 1. Ini mungkin disebabkan oleh teks *training* yang membahas banyak topik politik yang serupa dengan kalimat 2, sehingga memiliki banyak kata yang ada di keduanya. Ini juga didukung oleh akurasi kalimat 2 yang sudah lebih bagus dibandingkan kalimat 1 pada iterasi dengan vocab_size 500 pada kedua metode.

BPE v WordPiece

Akurasi pada metode BPE secara umum lebih bagus dibandingkan metode WordPiece pada kedua kalimat. Karena wordpiece memperhitungkan kemungkinan suatu kata untuk muncul, banyak singkatan dan angka di-*merge* terlebih dahulu untuk beberapa iterasi pertama. Ini karena kebanyakan kata yang memiliki huruf kapital di awal memiliki kesempatan besar untuk merupakan singkatan. Hal yang sama dapat dikatakan untuk tanggal di mana string dengan angka berkemungkinan tinggi merupakan tanggal. Bila kita melihat teks *training*, memang terdapat banyak singkatan dan angka sehingga kata-kata tersebut dimasukkan terlebih dahulu. Ini bisa juga dilihat dari hasil tokenisasi (PMA dan 2006 sudah ada pada iterasi dengan vocab_size 500).

BPE

BPE memiliki runtime yang jauh lebih singkat dan akurasi yang lebih tinggi untuk kedua kalimat yang diuji. Akurasi BPE sudah terlihat bagus pada vocab size 500 khususnya untuk kalimat 1. Meskipun begitu, ketika melihat hasil tokenisasi beberapa kata lain, BPE tidak dapat memproses clitic seperti -nya. Ini mungkin disebabkan oleh vocab size yang kurang tepat sehingga semua string berakhiran -nya tidak terpisah.

WordPiece

WordPiece memiliki runtime yang lebih lama dan tingkat akurasi yang buruk pada vocab size yang lebih kecil. Akurasi WordPiece terlihat khususnya buruk pada vocab size 500 iterasi-iterasi

pertama untuk kedua kalimat. Pada iterasi yang lebih besar, akurasi terlihat lebih bagus dengan 32 / 33 untuk kalimat 2 dengan vocab size 10000. WordPiece terlihat bagus ketika melakukan merge pada string yang lebih spesifik seperti pada singkatan dan angka. Kualitas lain dari wordpiece adalah kemampuannya untuk melakukan tokenisasi pada clitic.

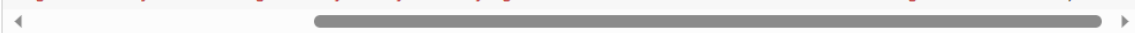
Program

BPE

vocab_size 500:

1

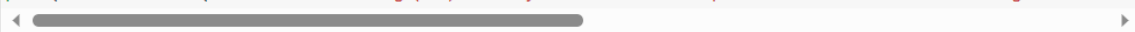
```
In [63]: mengerahkan monyet untuk mengusir monyet-monyet lain yang berbadan lebih kecil dari arena Pesta Olahraga Persemakmuran.').tokens;
```



```
['Pem', 'erintah', 'ko', 'ta', 'D', 'el', 'h', 'i', 'meng', 'er', 'ah', 'kan', 'm', 'on', 'y', 'et', 'untuk', 'meng', 'u', 's', 'i', 'r', 'm', 'on', 'y', 'et', '-', 'm', 'on', 'y', 'et', 'la', 'in', 'yang', 'ber', 'ba', 'dan', 'lebih', 'ke', 'c', 'il', 'da', 'ri', 'arena', 'P', 'es', 'ta', 'O', 'lah', 'r', 'aga', 'Per', 'sem', 'ak', 'm', 'un', 'an', '.']
```

2

```
In [10]: print(tokenizer.encode('Penanaman modal asing (PMA) di Malaysia tahun 2006 mencapai lima kali lebih besar dibandingkan Indonesia,
```

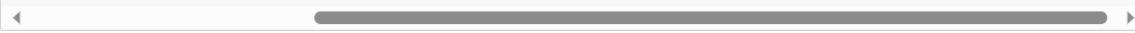


```
['Pen', 'an', 'am', 'an', 'modal', 'as', 'ing', '(', 'P', 'M', 'A', ')', 'di', 'M', 'al', 'ay', 'sia', 'tahun', '2006', 'mencap', 'ai', 'li', 'ma', 'k', 'ali', 'lebih', 'besar', 'dibanding', 'kan', 'Indonesia', ',', 'h', 'al', 'ini', 'men', 'un', 'j', 'uk', 'kan', 'pemb', 'ang', 'unan', 'ekonomi', 'M', 'al', 'ay', 'sia', 'ja', 'uh', 'lebih', 'men', 'ari', 'k', 'dibanding', 'kan', 'I', 'ndonesia', 'ba', 'gi', 'in', 'ves', 'tor', 'as', 'ing', '.']
```

vocab_size 1000:

1

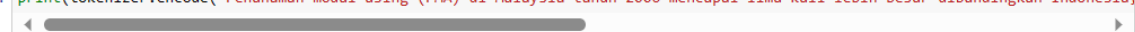
```
In [79]: mengerahkan monyet untuk mengusir monyet-monyet lain yang berbadan lebih kecil dari arena Pesta Olahraga Persemakmuran.').tokens;
```



```
['Pemerintah', 'ko', 'ta', 'D', 'el', 'hi', 'meng', 'erah', 'kan', 'm', 'ony', 'et', 'untuk', 'meng', 'u', 'si', 'r', 'm', 'on', 'y', 'et', '-', 'm', 'ony', 'et', 'lain', 'yang', 'ber', 'ba', 'dan', 'lebih', 'ke', 'c', 'il', 'dari', 'arena', 'P', 'es', 't', 'a', 'O', 'lah', 'r', 'aga', 'Per', 'sem', 'ak', 'm', 'uran', '.']
```

2

```
In [10]: print(tokenizer.encode('Penanaman modal asing (PMA) di Malaysia tahun 2006 mencapai lima kali lebih besar dibandingkan Indonesia,
```

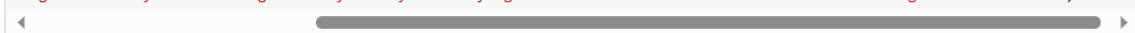


```
['Pen', 'an', 'aman', 'modal', 'asing', '(', 'P', 'M', 'A', ')', 'di', 'M', 'al', 'ay', 'sia', 'tahun', '2006', 'mencapai', 'li', 'ma', 'k', 'ali', 'lebih', 'besar', 'dibandingkan', 'Indonesia', ',', 'h', 'al', 'ini', 'menunjukkan', 'pembangunan', 'ekonomi', 'M', 'al', 'ay', 'sia', 'ja', 'uh', 'lebih', 'men', 'ari', 'k', 'dibandingkan', 'Indonesia', 'bagi', 'inves', 'tor', 'asing', '.']
```

vocab_size 5000:

1

```
In [87]: mengerahkan monyet untuk mengusir monyet-monyet lain yang berbadan lebih kecil dari arena Pesta Olahraga Persemakmuran.').tokens;
```



```
['Pemerintah', 'kota', 'Delhi', 'meng', 'erah', 'kan', 'monyet', 'untuk', 'mengusir', 'monyet', '-', 'monyet', 'lain', 'yang', 'berbadan', 'lebih', 'kecil', 'dari', 'arena', 'P', 'es', 'ta', 'O', 'lah', 'raga', 'Per', 'sem', 'ak', 'm', 'uran', '.']
```

2


```
In [10]: print(tokenizer.encode('Penanaman modal asing (PMA) di Malaysia tahun 2006 mencapai lima kali lebih besar dibandingkan Indonesia, hal ini menunjukkan pembangunan ekonomi Malaysia jauh lebih menarik dibandingkan Indonesia bagi investor asing.'))

['Pen', 'anaman', 'modal', 'asing', '(', 'PMA', ')', 'di', 'Malaysia', 'tahun', '2006', 'mencapai', 'lima', 'kali', 'lebih', 'b', 'esar', 'dibandingkan', 'Indonesia', ',', 'hal', 'ini', 'menunjukkan', 'pembangunan', 'ekonomi', 'Malaysia', 'jauh', 'lebih', 'm', 'enarik', 'dibandingkan', 'Indonesia', 'bagi', 'investor', 'asing', '.']
```

vocab_size 10000:

1

```
In [95]: mengerahkan monyet untuk mengusir monyet-monyet lain yang berbadan lebih kecil dari arena Pesta Olahraga Persemakmuran').tokens

['Pemerintah', 'kota', 'Delhi', 'meng', 'erah', 'kan', 'monyet', 'untuk', 'mengusir', 'monyet', '-', 'monyet', 'lain', 'yang', 'berbadan', 'lebih', 'kecil', 'dari', 'arena', 'P', 'es', 'ta', 'O', 'lah', 'raga', 'Pen', 'sem', 'ak', 'm', 'uran', '.']
```

2

```
In [10]: print(tokenizer.encode('Penanaman modal asing (PMA) di Malaysia tahun 2006 mencapai lima kali lebih besar dibandingkan Indonesia, hal ini menunjukkan pembangunan ekonomi Malaysia jauh lebih menarik dibandingkan Indonesia bagi investor asing.'))

['Penanaman', 'modal', 'asing', '(', 'PMA', ')', 'di', 'Malaysia', 'tahun', '2006', 'mencapai', 'lima', 'kali', 'lebih', 'besa', 'r', 'dibandingkan', 'Indonesia', ',', 'hal', 'ini', 'menunjukkan', 'pembangunan', 'ekonomi', 'Malaysia', 'jauh', 'lebih', 'mena', 'rik', 'dibandingkan', 'Indonesia', 'bagi', 'investor', 'asing', '.']
```

WordPiece

vocab_size 500:

1

```
In [15]: # Contoh tokenisasi dengan tokenizer yang sudah dilatih
print(tokenize('Pemerintah kota Delhi mengerahkan monyet untuk mengusir monyet-monyet lain yang berbadan lebih kecil dari arena Pesta Olahraga Persemakmuran').tokens)

['P', '##e', '##m', '##e', '##r', '##i', '##n', '##t', '##a', '##h', 'k', '##o', '##t', '##a', 'D', '##e', '##l', '##h', '##i', 'm', '##e', '##n', '##g', '##e', '##r', '##a', '##h', '##k', '##a', '##n', 'm', '##o', '##n', '##y', '##e', '##t', 'u', '##n', '##t', '##u', '##k', 'm', '##e', '##n', '##g', '##u', '##s', '##i', '##n', 'm', '##o', '##n', '##y', '##e', '##t', '-', 'm', '##o', '##n', '##y', '##e', '##t', 'l', '##a', '##i', '##n', 'y', '##a', '##n', '##g', 'b', '##e', '##n', '##b', '##a', '##d', '##a', '##n', 'l', '##e', '##b', '##i', '##h', 'k', '##e', '##c', '##i', '##l', 'd', '##a', '##r', '##l', 'a', '##r', '##e', '##n', '##a', 'P', '##e', '##s', '##t', '##a', 'O', '##l', '##a', '##h', '##r', '##a', '##g', '##a', 'P', '##e', '##r', '##s', '##e', '##m', '##a', '##k', '##m', '##u', '##r', '##a', '##n', '.']
```

2

```
In [16]: # Contoh melakukan tokenisasi suatu kalimat
print(tokenize('Penanaman modal asing (PMA) di Malaysia tahun 2006 mencapai lima kali lebih besar dibandingkan Indonesia, hal ini menunjukkan pembangunan ekonomi Malaysia jauh lebih menarik dibandingkan Indonesia bagi investor asing.'))

['P', '##e', '##n', '##a', '##n', '##a', '##m', '##a', '##n', 'm', '##o', '##d', '##a', '##l', 'a', '##s', '##i', '##n', '##g', '(', 'PMA', ')', 'd', '##i', 'M', '##a', '##l', '##a', '##y', '##s', '##i', '##a', 't', '##a', '##h', '##u', '##n', '2006', 'm', '##e', '##n', '##c', '##a', '##p', '##a', '##i', 'l', '##i', '##m', '##a', 'k', '##a', '##l', '##i', 'l', '##e', '##b', '##i', '##h', 'b', '##e', '##s', '##a', '##r', 'd', '##i', '##b', '##a', '##n', '##d', '##l', '##n', '##g', '##k', '##a', '##n', 'I', '##n', '##d', '##o', '##n', '##e', '##s', '##i', '##a', 'h', '##a', '##l', 'i', '##n', '##i', 'm', '##e', '##n', '##u', '##n', '##j', '##u', '##k', '##k', '##a', '##n', 'p', '##e', '##m', '##b', '##a', '##n', '##g', '##u', '##n', '##a', '##n', 'e', '##k', '##o', '##n', '##o', '##m', '##i', 'M', '##a', '##l', '##a', '##y', '##s', '##i', '##a', 'j', '##a', '##u', '##h', 'l', '##e', '##b', '##i', '##h', 'm', '##e', '##n', '##a', '##r', '##l', '##k', 'd', '##l', '##b', '##a', '##n', '##d', '##i', '##n', '##g', '##k', '##a', '##n', 'I', '##n', '##d', '##o', '##n', '##e', '##s', '##i', '##a', 'b', '##a', '##g', '##i', 'i', '##n', '##v', '##e', '##s', '##t', '##o', '##r', 'a', '##s', '##i', '##n', '##g', '.']
```

vocab_size 1000:

1

```
In [15]: # Contoh tokenisasi dengan tokenizer yang sudah dilatih
print(tokenize('Pemerintah kota Delhi mengerahkan monyet untuk mengusir monyet-monyet lain yang berbadan lebih kecil dari arena f

['P', '##e', '##m', '##e', '##r', '##i', '##n', '##t', '##a', '##h', 'k', '##o', '##t', '##a', 'D', '##e', '##l', '##h', '##i',
'm', '##e', '##n', '##g', '##e', '##r', '##a', '##h', '##k', '##a', '##n', 'm', '##o', '##n', '##y', '##e', '##t', 'u', '##n',
'##t', '##u', '##k', 'm', '##e', '##n', '##g', '##u', '##s', '##i', '##r', 'm', '##o', '##n', '##y', '##e', '##t', '-', 'm', '##
o', '##n', '##y', '##e', '##t', 'l', '##a', '##i', '##n', 'y', '##a', '##n', '##g', 'b', '##e', '##r', '##b', '##a', '##d', '#
#a', '##n', 'l', '##e', '##b', '##i', '##h', 'k', '##e', '##c', '##i', '##l', 'd', '##a', '##r', '##i', 'a', '##r', '##e', '##
n', '##a', 'P', '##e', '##s', '##t', '##a', 'O', '##l', '##a', '##h', '##r', '##a', '##g', '##a', 'P', '##e', '##r', '##s', '##
e', '##m', '##a', '##k', '##m', '##u', '##r', '##a', '##n', '##.']
```

2

```
In [16]: # Contoh melakukan tokenisasi suatu kalimat
print(tokenize('Penanaman modal asing (PMA) di Malaysia tahun 2006 mencapai lima kali lebih besar dibandingkan Indonesia, hal ini

['P', '##e', '##n', '##a', '##n', '##a', '##m', '##a', '##n', 'm', '##o', '##d', '##a', '##l', 'a', '##s', '##i', '##n', '##g',
('PMA', 't'), 'd', '##i', 'M', '##a', '##l', '##a', '##y', '##s', '##i', '##a', 't', '##a', '##h', '##u', '##n', '2006',
'm', '##e', '##n', '##c', '##a', '##p', '##a', '##i', 'l', '##i', '##m', '##a', '##i', 'l', '##e', '##b', '#
#l', '##h', 'b', '##e', '##s', '##a', '##r', 'd', '##i', '##b', '##a', '##n', '##d', '##l', '##n', '##g', '##k', '##a', '##n',
'I', '##n', '##d', '##o', '##n', '##e', '##s', '##i', '##a', 'h', '##a', '##l', 'i', '##n', '##i', 'm', '##e', '##n', '##
u', '##n', '##j', '##u', '##k', '##a', '##n', 'p', '##e', '##m', '##b', '##a', '##n', '##g', '##u', '##n', '##a', '##n', '##
e', '##k', '##o', '##n', '##o', '##m', '##i', 'M', '##a', '##l', '##a', '##y', '##s', '##l', '##a', 'j', '##a', '##u', '##h',
'l', '##e', '##b', '##i', '##h', 'm', '##e', '##n', '##a', '##n', '##i', '##k', 'd', '##i', '##b', '##a', '##n', '##d', '##i',
'##n', '##g', '##k', '##a', '##n', 'I', '##n', '##d', '##o', '##n', '##e', '##s', '##i', '##a', 'b', '##a', '##g', '##l', 'i',
'##n', '##v', '##e', '##s', '##t', '##o', '##r', 'a', '##s', '##l', '##n', '##g', '##.']
```

vocab_size 5000:

1

```
In [15]: # Contoh tokenisasi dengan tokenizer yang sudah dilatih
print(tokenize('Pemerintah kota Delhi mengerahkan monyet untuk mengusir monyet-monyet lain yang berbadan lebih kecil dari arena f

['Pemerint', '##a', '##h', 'kot', '##a', 'Delhi', 'menger', '##a', '##h', '##k', '##a', '##n', 'monyet', 'untuk', 'mengu', '##
s', '##i', '##r', 'monyet', '-', 'monyet', 'l', '##a', '##i', '##n', 'y', '##a', '##ng', 'berb', '##a', '##d', '##a', '##n', 'l
ebih', 'kecil', 'd', '##a', '##r', '##i', 'ar', '##e', '##n', '##a', 'Pe', '##s', '##t', '##a', 'Ol', '##a', '##h', '##r', '##
a', '##g', '##a', 'Perse', '##m', '##a', '##k', '##m', '##u', '##r', '##a', '##n', '##.']
```

2

```
In [16]: # Contoh melakukan tokenisasi suatu kalimat
print(tokenize('Penanaman modal asing (PMA) di Malaysia tahun 2006 mencapai lima kali lebih besar dibandingkan Indonesia, hal ini

['Pen', '##a', '##n', '##a', '##m', '##a', '##n', 'mod', '##a', '##l', 'asing', '(', 'PMA', 't'), 'd', '##i', 'M', '##a', '##l',
'##a', '##y', '##s', '##i', '##a', 't', '##a', '##h', '##u', '##n', '2006', 'menc', '##a', '##p', '##a', '##i', 'l', '##i', '##
m', '##a', 'k', '##a', '##l', '##i', 'lebih', 'bes', '##a', '##r', 'd', '##i', '##b', '##a', '##n', '##d', '##i', '##ng', '##
k', '##a', '##n', 'Indonesia', '##a', 'h', '##a', '##l', 'ini', 'menunjukk', '##a', '##n', 'pemb', '##a', '##ng', '##u', '#
n', '##a', '##n', 'ekonomi', 'M', '##a', '##l', '##a', '##y', '##s', '##i', '##a', 'j', '##a', '##u', '##h', 'lebih', 'men',
'##a', '##r', '##i', '##k', 'd', '##i', '##b', '##a', '##n', '##d', '##i', '##ng', '##k', '##a', '##n', 'Indonesia', '##a', 'b',
'##a', '##g', '##i', 'investor', 'asing', '##.']
```

vocab_size 10000:

1

```
In [15]: # Contoh tokenisasi dengan tokenizer yang sudah dilatih
print(tokenize('Pemerintah kota Delhi mengerahkan monyet untuk mengusir monyet-monyet lain yang berbadan lebih kecil dari arena f

['Pemerintah', 'kot', '##a', 'Delhi', 'menger', '##ahk', '##a', '##n', 'monyet', 'untuk', 'mengusir', 'monyet', '-', 'monyet',
'lain', 'yang', 'berbadan', 'lebih', 'kecil', 'd', '##a', '##r', 'an', '##e', '##n', '##a', 'Pe', '##s', '##t', '##a', 'Ol',
'##a', '##h', '##r', '##a', '##ga', 'Perse', '##m', '##a', '##k', '##mu', '##r', '##a', '##n', '##.']
```

2

```
In [16]: # Contoh melakukan tokenisasi suatu kalimat
print(tokenize('Penanaman modal asing (PMA) di Malaysia tahun 2006 mencapai lima kali lebih besar dibandingkan Indonesia, hal ini

['Pen', '##a', '##n', '##a', '##m', '##a', '##n', 'modal', 'asing', '(', 'PMA', ')', 'di', 'Malaysia', 'tahun', '2006', 'mencap
ai', 'lima', 'kali', 'lebih', 'besar', 'dibandingkan', 'Indonesia', ',', 'hal', 'ini', 'menunjukkan', 'pembangunan', 'ekonomi',
'Malaysia', 'jauh', 'lebih', 'menarik', 'dibandingkan', 'Indonesia', 'bagi', 'investor', 'asing', '.']
```

Kesimpulan

1. Teks *training* lebih cocok untuk digunakan pada teks yang berhubungan dengan politik
2. BPE lebih bagus dibandingkan WordPiece pada kedua kalimat karena hasil tokenisasi WordPiece memprioritaskan banyak singkatan dan angka sehingga memerlukan banyak iterasi untuk bisa bagus.
3. Terdapat beberapa kata yang tidak ada di dalam tokenizer seperti 'olahraga' dan 'persemakmuran'.

2. Algoritma

Program untuk membaca teks

```
train_text = ""
gold_standard = []

with open('test.txt', 'r') as test:
    lines = test.readlines()
    for line in lines:
        if line.startswith('# text ='):
            train_text += line[9:]
        else:
            gold_standard.append(line.strip())

# print(train_text)
# print(gold_standard)
```

Program untuk menghitung akurasi BPE

```
In [15]: def tokenizer_test(tokenizer, gold_standard, text):
    accurate_tokenized = 0
    total_corpus = len(gold_standard)
    gs_copy = gold_standard.copy()
    current_number = 0

    tokenize_result = tokenizer.encode(text).tokens

    for word in tokenize_result:
        if word in gs_copy[current_number:total_corpus]:
            accurate_tokenized += 1
            for num in range(current_number, len(gs_copy)):
                if gs_copy[num] != word:
                    current_number += 1
                    break

    print(str(accurate_tokenized) + " / " + str(total_corpus))
    return accurate_tokenized / total_corpus

print(tokenizer_test(tokenizer, gold_standard, train_text))

9092 / 10041
0.905487501244896
```

Program untuk menghitung akurasi WordPiece

```
def tokenizer_test(gold_standard, text):
    accurate_tokenized = 0
    total_corpus = len(gold_standard)
    gs_copy = gold_standard.copy()
    current_number = 0

    tokenize_result = tokenize(text)

    for word in tokenize_result:
        word_proc = word

        if word_proc[:2] == "##":
            word_proc = word_proc[2:]

        if word_proc in gs_copy[current_number:total_corpus]:
            accurate_tokenized += 1
            for num in range(current_number, len(gs_copy)):
                if gs_copy[num] != word:
                    current_number += 1
                    break
            if len(gs_copy) == 0:
                break

    print(str(accurate_tokenized) + " / " + str(total_corpus))
    return accurate_tokenized / total_corpus

print(tokenizer_test(gold_standard, train_text))
```

Cara kerja program adalah sebagai berikut.

1. Program membuka file 'test.txt' dan membacanya. Setiap line yang memiliki '#text=' ditambahkan kedalam string 'train_text' sementara line lain akan dimasukkan ke dalam list 'gold_standard'
2. Setelah itu program melakukan tokenisasi pada 'train_text' untuk mendapatkan hasil tokenisasi
3. Dalam method 'tokenizer_text', buat beberapa variabel untuk membantu perbandingan.
 - a. 'Accurate_tokenized' untuk menghitung berapa hasil tokenisasi yang akurat
 - b. 'Total_corpus' sebagai panjang gold standard
 - c. 'Current_number' untuk menghitung posisi pointer dalam array

4. Perbandingan dilakukan sebagai berikut:
 - a. Iterasikan untuk setiap kata yang ada di gold_standard. Kata untuk setiap iterasi adalah 'word'
 - b. Bila 'word' ada di dalam gold_standard dari index 'current_number' sampai akhir, tambahkan 'accurate_tokenized' untuk menandakan adanya kata di dalam korpus
 - c. Dari 'current_number' sampai akhir gold_standard, iterasikan hingga ketemu 'word'. Untuk setiap kata yang bukan 'word' tambahkan 'current_number' dengan satu. Ini memastikan tidak terjadinya perulangan
 - d. Lakukan hingga selesai

Ada satu langkah tambahan untuk menghitung akurasi WordPiece yaitu untuk menghapus semua string yang memiliki "##" di awal.