

28 APRIL 2024



Assignment Chapter 5

INTRODUCTION TO ARTIFICIAL INTELLIGENCE

PRESENTED BY

Tim 1A - OneOnlyOne

[Link notebook](#)

Member

SC60119	Arofiqul Aufa Diennoor Alwitri
SC60117	Niken Ayuputri
SC60123	Muhammad Maulidinal Haqi Al Azizi
SC60112	Adam Raihan Deha
SC60105	Geytsa Yumna Rabbania Rahman
SC60103	Cristy Elisabeth Putri Bojoh
SC60106	Khoirunnisa

Agenda

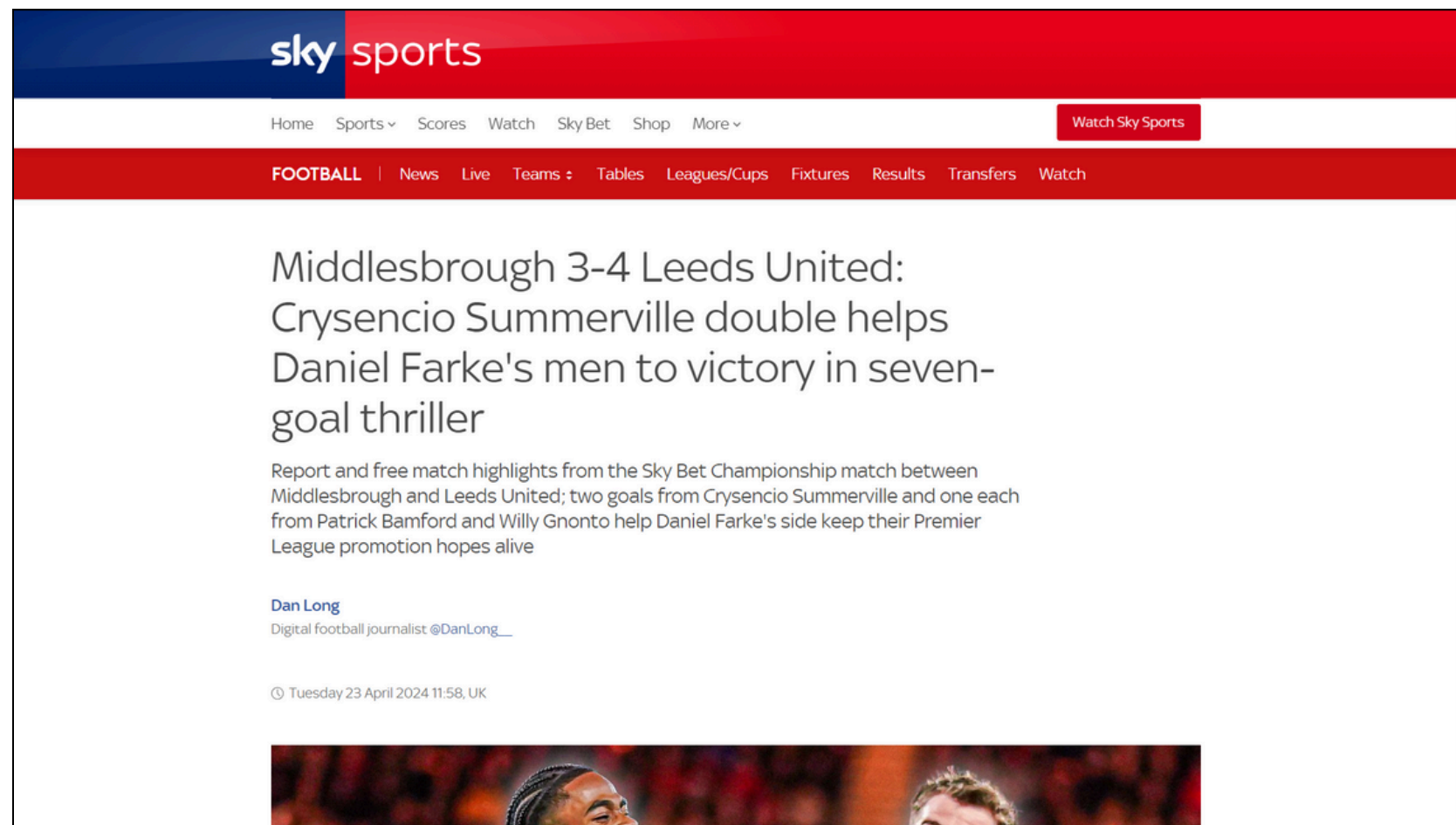
- 3 Introduction
- 7 Implementation

Introduction



Artikel/ Teks

Pada implementasi dari Artificial Intelligence ditentukan **tiga artikel atau teks** sebagai dasar dengan topik: **olahraga (sport)**, **kesehatan (medical)**, dan **keuangan (finance)**. Artikel atau teks yang digunakan adalah sebagai berikut.



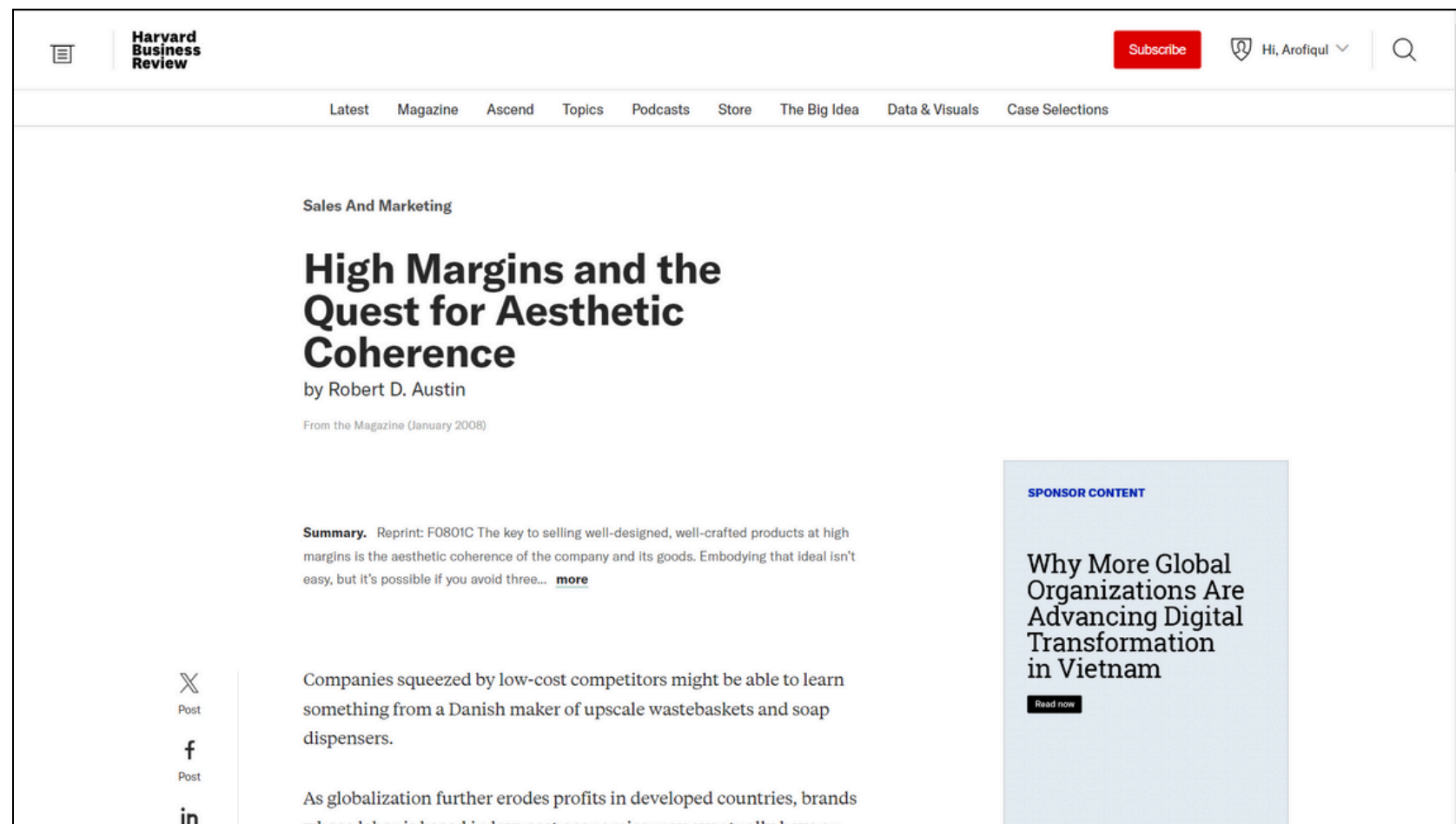
artikel olahraga



artikel kesehatan

Artikel/ Teks

Pada implementasi dari Artificial Intelligence ditentukan [tiga artikel atau teks](#) sebagai dasar dengan topik: [olahraga \(sport\)](#), [kesehatan \(medical\)](#), dan [keuangan \(finance\)](#). Artikel atau teks yang digunakan adalah sebagai berikut.



artikel keuangan

Implementation



Text Cleaning

Ketiga artikel atau teks yang digunakan dilakukan **pengolahan dengan awal**, yaitu **tokenization** untuk membagi seluruh teks menjadi token, menghilangkan dan **membersihkan teks dari stopwords** atau kata yang tidak menambah banyak makna pada sebuah kalimat seperti kata ganti orang dan kata tanya, **lematization atau stemming** untuk mengembalikan **kata kunci penting** dalam teks.

Dalam ketiga teks, secara singkat didapatkan hasil sebagai berikut.

```
{'hope', 'offside', 'per', 'WHO', 'considered', 'man', 'equivalent', 'extend', 'profit', 'digit', 'Every', 'back', "n't", 'create',
```

kata kunci penting dalam ketiga teks

```
["Leeds maintained automatic promotion charge thrilling 4-3 win Middlesbrough Riverside Stadium . Had Leeds lost , doo
```

teks setelah dilakukan pembersihan

Text Cleaning

Selain ketiga topik artikel atau teks tersebut, **teks yang akan dilakukan pengklasifikasian** juga **dilakukan pengolahan data awal** dengan **tokenization** untuk membagi seluruh teks menjadi token, menghilangkan dan **membersihkan teks dari stopwords** atau kata yang tidak menambah banyak makna pada sebuah kalimat seperti kata ganti orang dan kata tanya, **lematization atau stemming** untuk mengembalikan kata kunci penting dalam teks.

Teks yang digunakan untuk klasifikasi adalah sebagai berikut.

"Cristiano Ronaldo came off the bench to earn Manchester United a hard-fought 2-1 victory at Everton in the Premier League on Sunday, taking his career goal tally to 700 in the process. Just as United did last weekend in their derby mauling at the hands of local rivals Manchester City, they again found themselves behind early on at Goodison Park after Alex Iwobi curled a sublime strike into the net from 20 metres."

teks yang akan diklasifikasikan

Bag of Words

Selanjutnya, setelah seluruh teks dibersihkan dari stopwords, dilakukan proses lematization atau stemming maka akan dihitung **frekuensi untuk setiap kata dalam ketiga topik** dengan **menghitung kemunculan setiap kata**. Kemudian berdasarkan kemunculannya, ditambahkan **nilai 0 atau 1** sesuai dengan kemunculannya dalam kalimat. Penambahan **nilai 0 menunjukkan bahwa kata tersebut tidak terdapat dalam teks**, sedangkan penambahan **nilai 1 menunjukkan adanya kata** tersebut dalam teks.

Hasil perhitungan tersebut akan **dipetakan** dalam **vektor bag-of-words** yang akan dipetakan untuk **setiap topik teks**. Sedangkan, **teks klasifikasi yang telah bersih** dilakukan **pemetaan** kata ke dalam **hasil bag-of-words dari ketiga topik** tersebut untuk mengetahui pada **topik** apa **teks** yang akan diklasifikasikan memiliki kecenderungan kemiripan.

Bag of Words

Hasil pemetaan vektor bag-of-words untuk setiap topik teks dan teks yang akan diklasifikasikan menunjukkan terdapat total 724 kata dalam bag-of words untuk ketiga topik teks. Selanjutnya setelah vektor diubah menjadi array, didapatkan array dengan 4 baris yang setiap barisnya mewakili teks yang dipetakan, yaitu teks sport, medical, finance, dan teks yang akan diklasifikasikan.

```
[[0 1 0 ... 2 0 2]
 [1 1 1 ... 0 3 0]
 [0 0 0 ... 1 0 0]
 [0 0 0 ... 0 0 0]]
Bag-of-Words vector count: 724
```

vektor bag-of words

	000	10	11	12	13	15th	17	1939	1957	1960s	...	with	within	wondered	wore	workmanship	world	worldwide	would	year	yet
text_sport	0	1	0	0	0	0	0	0	0	0	...	1	0	0	0	0	0	0	2	0	2
text_medical	1	1	1	1	1	1	2	0	0	0	...	0	1	0	0	0	2	3	0	3	0
text_finance	0	0	0	0	0	0	0	1	1	1	...	0	0	1	1	2	2	0	1	0	0
query_text	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0

4 rows x 724 columns

array dari vektor bag-of-words

Klasifikasi

Hasil pemetaan bag-of-words pada teks yang akan diklasifikasikan akan dibandingkan dengan hasil pemetaan bag-of-words pada tiga topik teks: olahraga, kesehatan, dan keuangan untuk mengetahui kecondongan kemiripan pada salah satu topik teks.

Dilakukan normalisasi dari hasil vektor bag-of-words dan dilakukan perhitungan kemiripan dengan dot product (perkalian titik), selanjutnya akan dicari nilai similarity tertinggi diantara ketiga topik. Apabila nilai similarity terbesar adalah vektor index ke 0 maka teks diklasifikasikan sebagai teks sport, index ke 1 diklasifikasikan sebagai teks medical, dan terakhir atau index ke 2 diklasifikasikan sebagai teks finance.

Berdasarkan hasil didapatkan bahwa similarity terbesar berada pada index 0 dengan nilai maksimal 0.11535, sehingga teks diklasifikasikan sebagai teks olahraga (sport).

max *min* *in between*
[0.11535262482477, 0.0, 0.0324999238283928]
The query text is classified as: Sport

hasil klasifikasi

Terima Kasih

LINK NOTEBOOK:

[Assignment 5 - Introduction
to Artificial Intelligence](#)

