Національний Технічний Університет України «Київський Політехнічний Інститут»
Навчально-науковий комплекс «Інститут прикладного системного аналізу»
Кафедра математичних методів системного аналізу

Дипломна робота
освітньо-кваліфікаційного рівня «Бакалавр»
з напряму підготовки: «Системний аналіз»

на тему: Аналіз дописів соціальної мережі “Twitter” для визначення рейтингів футбольних команд

Виконав:
студент IV курсу
групи КА-21
Солопатич Р.І.

Науковий керівник:
к.т.н. доц. Дідковська М.В.

Мета, предмет та об’єкт дослідження

- **Мета дослідження –** розробити підхід до визначення рейтингів футбольних команд за популярністю в соцмережі Twitter на основі класифікатора Naïve Bayes, модель якого навчається на автоматично сформованих вибірках даних
- **Об’єкт дослідження** являє собою соціальну мережу Twitter, як великий масив текстів на різну тематику доступних до вільного доступу та аналізу.
- **Предметом дослідження** слугують методи автоматизованого аналізу тексту зокрема статистична модель семантичної пам’яті LSA (Latent Semantic Analysis) та алгоритм машинного навчання для класифікації Naïve Bayes.

Актуальність роботи

- Прослідковуючи за твітами Елона Маска (Elon Musk) – головного виконавчого директора компанії Tesla, було встановлено зв’язок між їх інтенсивністю та зміною цін акцій компанії Tesla
- Спорт став рекламним бізнесом. Популярність футбольної команди є її основною рисою, яка приваблює спонсорів та приносить кошти.
- Інтенсивності обговорення команд в соцмережі відображають їх популярність, а їх порівняння визначає *рейтинги популярності* команд в соцмережі

Задачі дипломного проекту

- Проаналізувати соціальну мережу Twitter і сформувати рейтинги футбольних команд за популярністю (інтенсивністю обговорення) в цій соцмережі, реалізувавши систему аналізу текстів твітів

- Розробити алгоритм кластеризації текстів дописів соцмережі Twitter та автоматизованого формування тренувальної вибірки моделі алгоритму машинного навчання

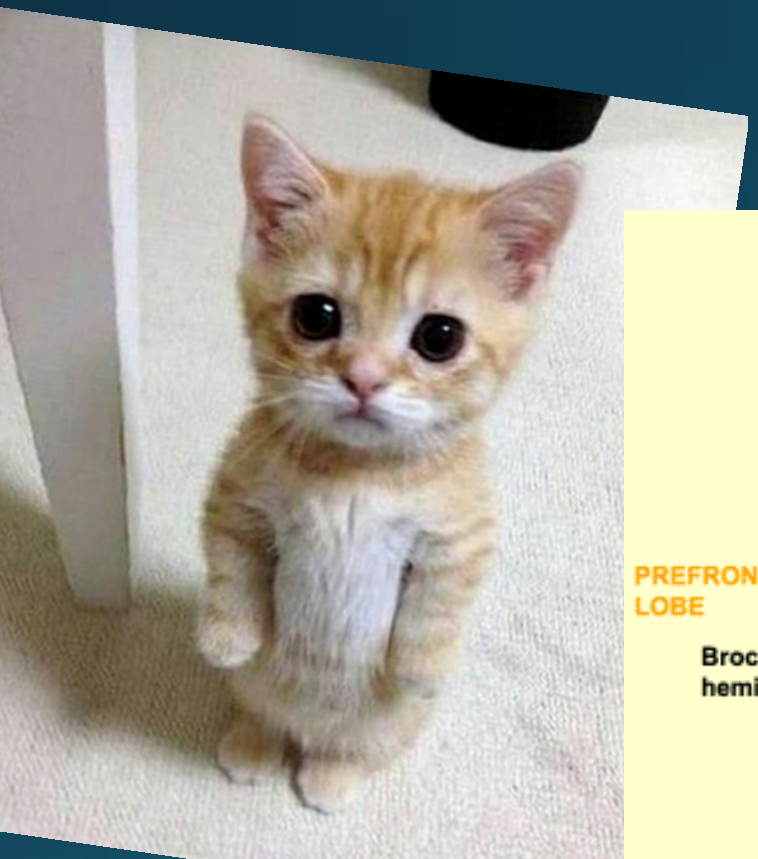
Існуючі підходи до розв’язання задачі класифікації текстів

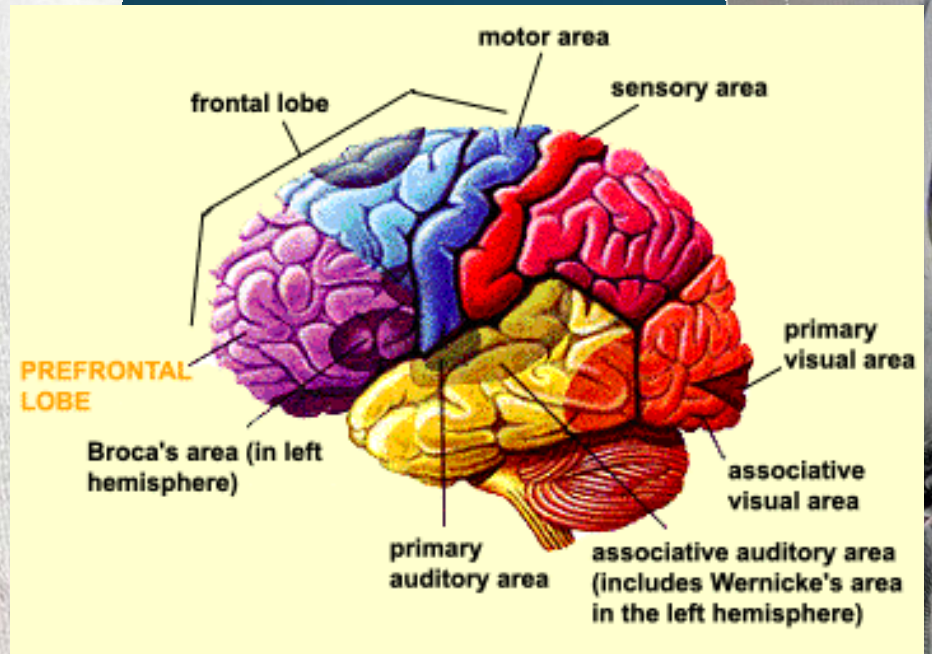
- Логістична регресія LogReg
- Multinomial Naïve Bayes
- SVM (Support Vector Machine) with linear kernel
- SVM (Support Vector Machine) with Gaussian kernel
- Naïve Bayes SVM (Support Vector Machine)

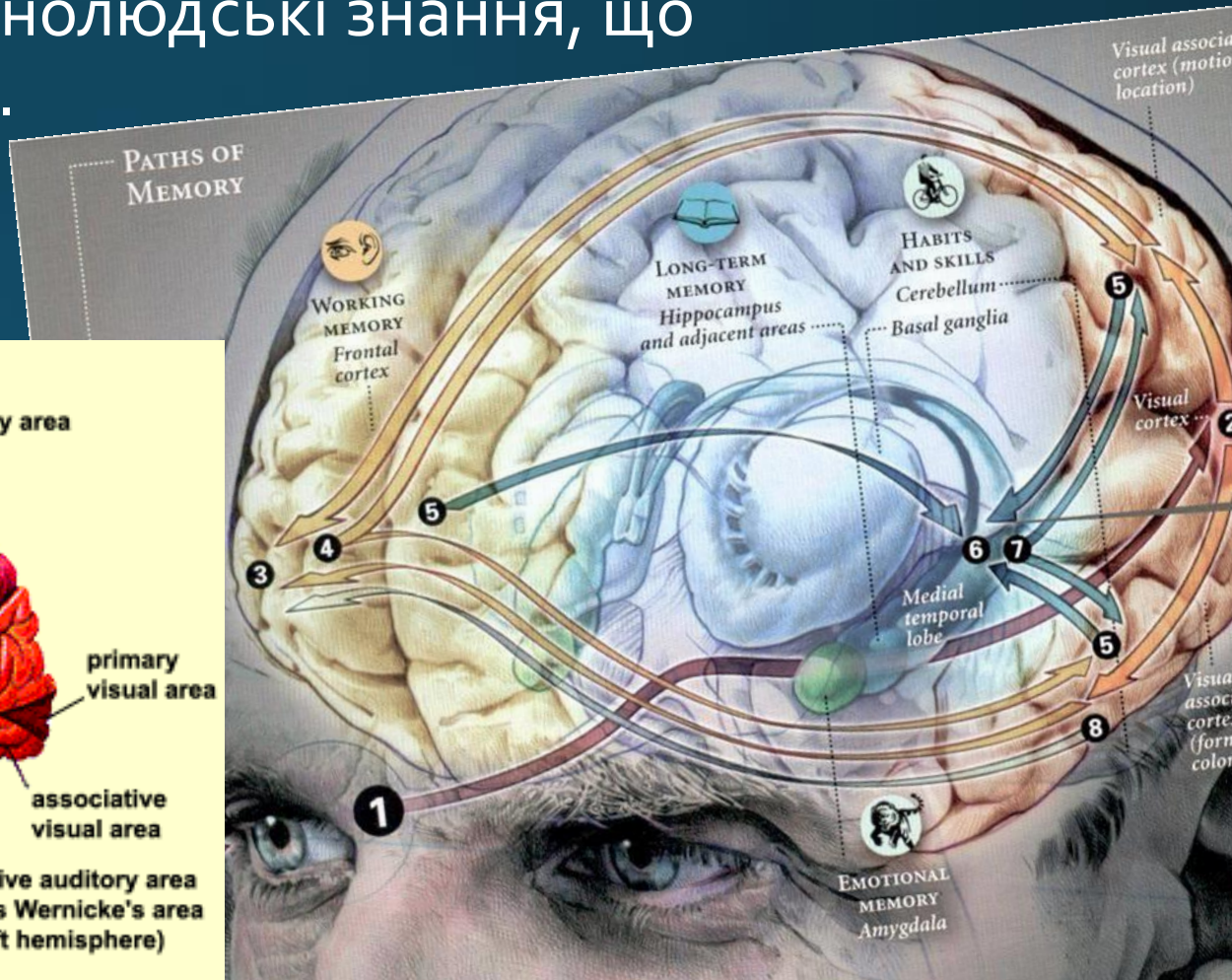
Основний недолік методів – необхідність наявності вибірки для тренування моделі, яка використовується у подальшому прогнозуванні

Математичний апарат дослідження

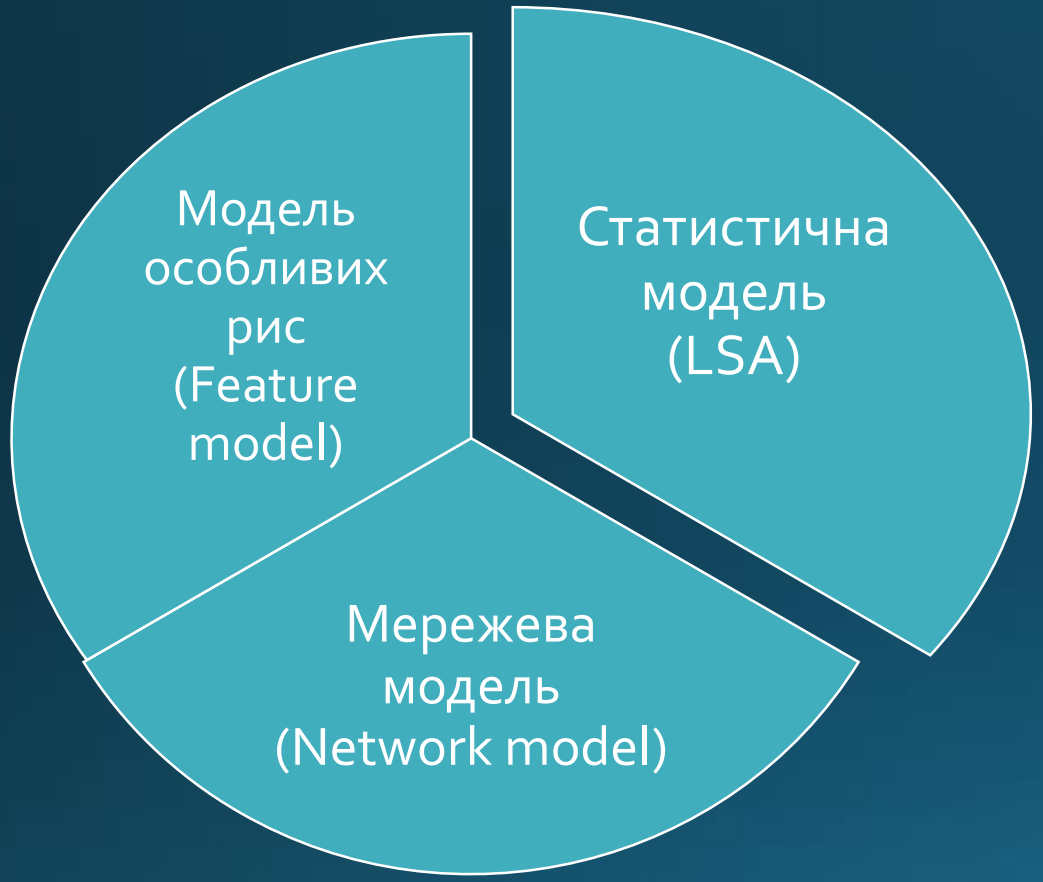
- Семантична пам’ять – людська пам’ять, яка явно зберігає факти чи події посилаючись на загальнолюдські знання, що акумулюються протягом життя.







Моделі семантичної пам’яті



Latent Semantic Analysis

Переваги

Недоліки

не потребує специфічних структур даних

повністю обслуговується математичним апаратом статистики та лінійної алгебри

ресурсоємність обчислень

Модель LSA

Example of text data: Titles of Some Technical Memos

c1: *Human* machine *interface* for ABC *computer* applications
c2: A *survey* of *user* opinion of *computer system response time*
c3: The *EPS user interface* management *system*
c4: *System* and *human system* engineering testing of *EPS*
c5: Relation of *user* perceived *response time* to error measurement

m1: The generation of random, binary, ordered *trees*
m2: The intersection *graph* of paths in *trees*
m3: *Graph minors* IV: Widths of *trees* and well-quasi-ordering
m4: *Graph minors*: A *survey*

$$\cos(v_1 \wedge v_2) = \frac{(v_1, v_2)}{\|v_1\|\|v_2\|} = \frac{(v_1, v_2)}{\sqrt{(v_1, v_1)}\sqrt{(v_2, v_2)}}$$

	c1	c2	c3	c4	c5	m1	m2	m3	m4
human	1	0	0	1	0	0	0	0	0
interface	1	0	1	0	0	0	0	0	0
computer	1	1	0	0	0	0	0	0	0
user	0	1	1	0	1	0	0	0	0
system	0	1	1	2	0	0	0	0	0
response	0	1	0	0	1	0	0	0	0
time	0	1	0	0	1	0	0	0	0
EPS	0	0	1	1	0	0	0	0	0
survey	0	1	0	0	0	0	0	0	1
trees	0	0	0	0	0	1	1	1	0
graph	0	0	0	0	0	0	1	1	1
minors	0	0	0	0	0	0	0	1	1

	c1	c2	c3	c4	c5	m1	m2	m3	m4
human	0.16	0.40	0.38	0.47	0.18	-0.05	-0.12	-0.16	-0.09
interface	0.14	0.37	0.33	0.40	0.16	-0.03	-0.07	-0.10	-0.04
computer	0.15	0.51	0.36	0.41	0.24	0.02	0.06	0.09	0.12
user	0.26	0.84	0.61	0.70	0.39	0.03	0.08	0.12	0.19
system	0.45	1.23	1.05	1.27	0.56	-0.07	-0.15	-0.21	-0.05
response	0.16	0.58	0.38	0.42	0.28	0.06	0.13	0.19	0.22
time	0.16	0.58	0.38	0.42	0.28	0.06	0.13	0.19	0.22
EPS	0.22	0.55	0.51	0.63	0.24	-0.07	-0.14	-0.20	-0.11
survey	0.10	0.53	0.23	0.21	0.27	0.14	0.31	0.44	0.42
trees	-0.06	0.23	-0.14	-0.27	0.14	0.24	0.55	0.77	0.66
graph	-0.06	0.34	-0.15	-0.30	0.20	0.31	0.69	0.98	0.85
minors	-0.04	0.25	-0.10	-0.21	0.15	0.22	0.50	0.71	0.62

Вхідні дані та очікувані результати

Назви команд

Початкові контексти

Ключові слова контекстів

Згруповані твіти по футбольних командах

Рейтинги команд за кількістю твітів в групах

Схема алгоритму кластеризації текстів та визначення рейтингів футбольрних команд

Кластери

- Формування за допомогою LSA початкових кластерів

Вибірка

- Завантаження нових твітів та визначення їх належності до початкових кластерів за допомогою LSA
- Формування тренувальної вибірки з нових кластеризованих дописів

Рейтинги

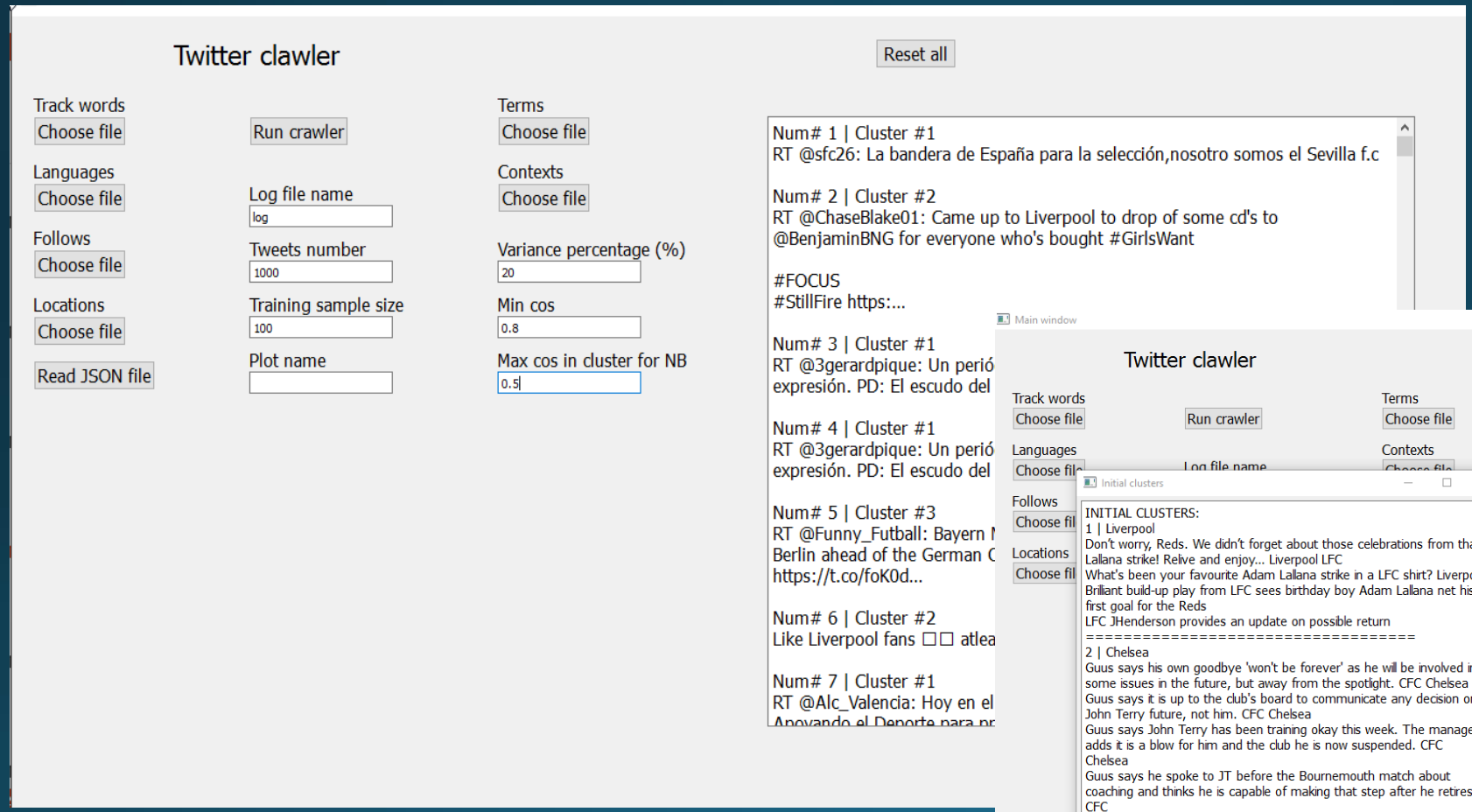
- Тренування класифікатора Naïve Bayes
- Аналіз великої вибірки нових завантажених твітів класифікатором
- Обчислення рейтингів команд на основі кількості оброблених дописів, які відносяться до кожної з команд

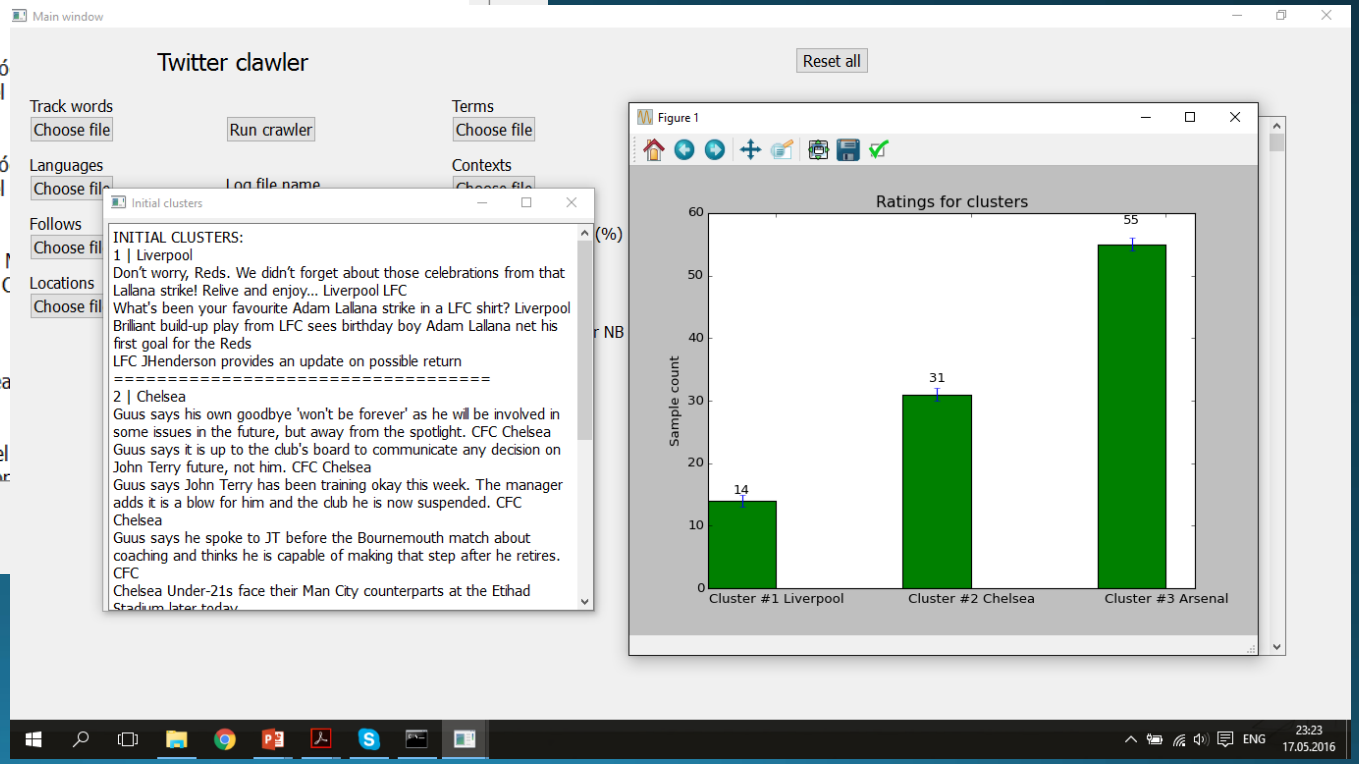
Алгоритми кластеризації та класифікації текстів


```
X:                                  Y:
[1,0,0,1,0,0,0,0,0,0,0,0]    --->   [2]
[0,0,0,1,0,0,0,0,0,1,0,0]    --->   [0]
[0,0,1,0,0,0,0,2,0,0,0]      --->   [1]
.......................              ...

[1,0,0,0,0,0,0,3,0,0,0]      --->   [2]
```

Інтерфейс користувача

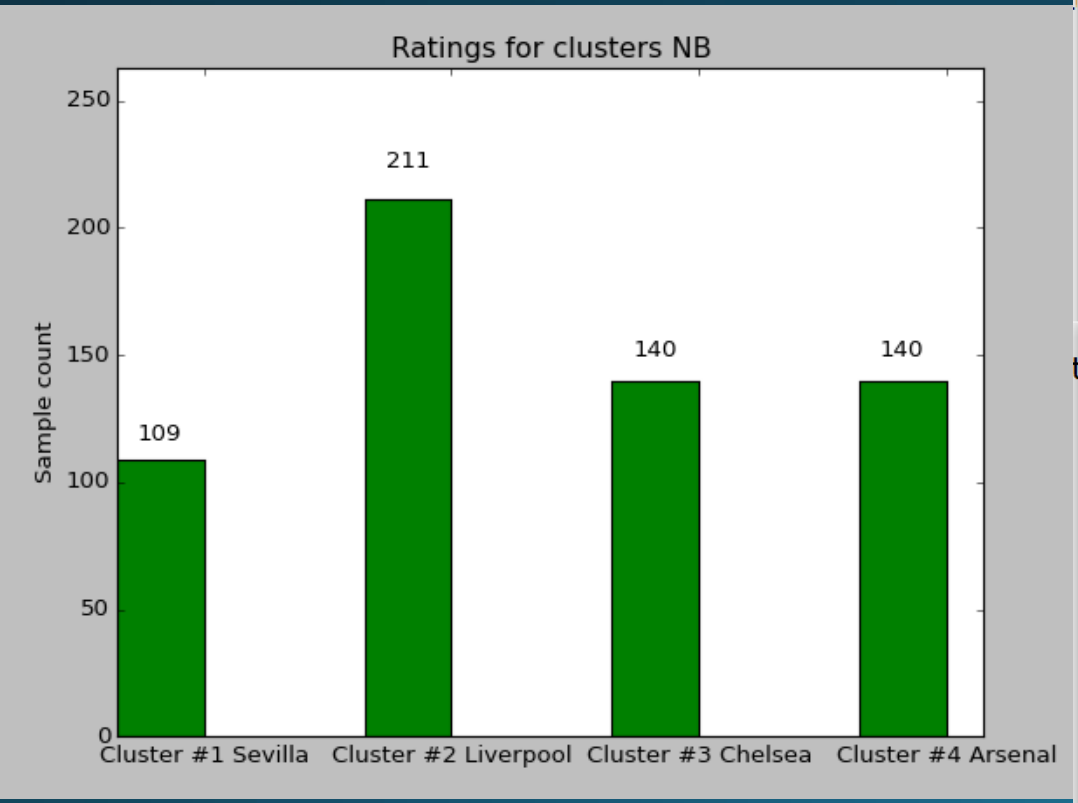




Визначення рейтингів команд

Визначення рейтингів команд (Sevillia, Liverpool, Chelsea, Arsenal)

Вхідні дані: об’єм навчальної вибірки 100, к-ть 500, подібність LSA 0,85, подібність NB 0,44 дисперсія 30% (24 травня 2016)



INITIAL CLUSTERS:
Cluster# 1 | Sevilla
El sentimentalismo llego al comienzo, con un video en recuerdo de Antonio Puerta, cuyo gol supuso el comienzo de la leyenda europea del SevillaFC
El SevillaFC repite con el premio Real Alcazar
SevillaFC putting the pressure on @LFC, both managers feeling tense vamosmisevilla APorLaQuinta Sevilla
First goal attempt by LFC blocked by @DavidSoria18 SevillaFC vamosmisevilla APorLaQuinta
Goal by @LFC, Sturridge in minute 34. Liverpool FC 1-0 #SevillaFC #vamosmisevilla #APorLaQuinta Sevilla
The arrival of SevillaFC at St. Jakob Park APorLaQuinta vamosmisevilla
==========
Cluster# 2 | Liverpool
Don’t worry, Reds. We didn’t forget about those celebrations from that Lallana strike! Relive and enjoy... LiverpoolFC LFC

Initial clusters

==========
Cluster# 3 | Chelsea
Guus says his own goodbye 'won't be forever' as he will be involved in some issues in the future, but away from the spotlight. CFC ChelseaFC
Guus says it is up to the club's board to communicate any decision on John Terry future, not him. CFC Chelsea
Guus says John Terry has been training okay this week. The manager adds it is a blow for him and the club he is now suspended. CFC Chelsea
Guus says he spoke to JT before the Bournemouth match about coaching and thinks he is capable of making that step after he retires. CFC
Chelsea Under-21s face their Man City counterparts at the Etihad Stadium later today ChelseaFC
Guus Hiddink reflects on several factors in the weekend defeat at Sunderland
==========
Cluster# 4 | Arsenal
IsaacHayden65 is on the Matchday Show for Arsenal Under-21s' play-off final. Get your questions in for him now
ANALYSIS: Get a look at the key tactics and stats from MCFCv AFC in The Breakdown. ArsenalFC

Reset all

NOW on Facebook Live: ... https://t.co/FUCLC9msUe

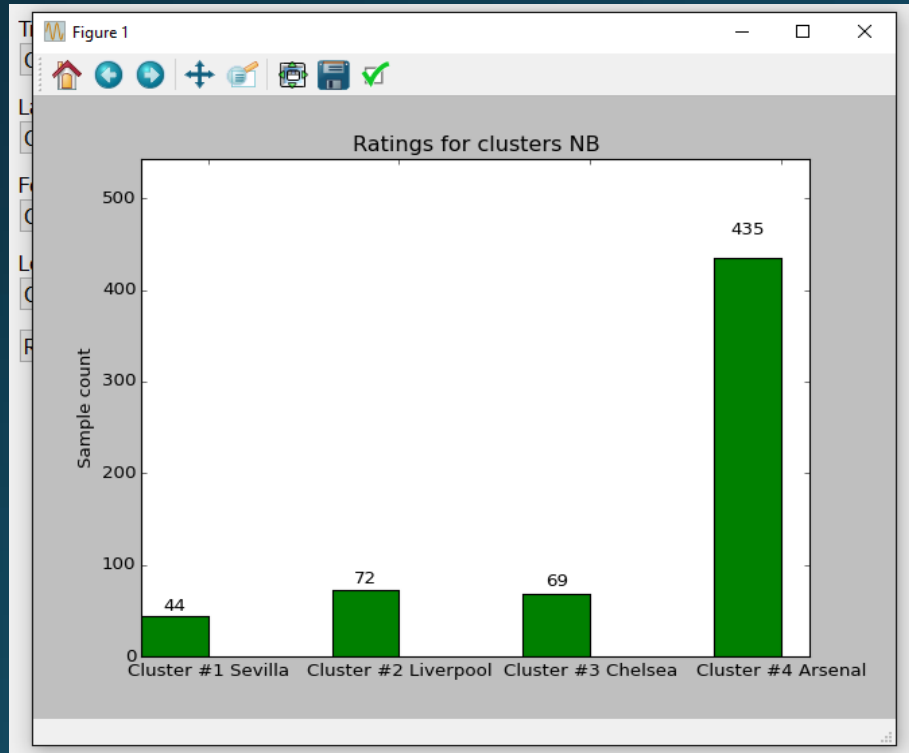
Num# 173 | Cluster #3
RT @ChelsTransfer: Chelsea have identified #Inter goalkeeper Samir Handanovic as their preferred goalkeeping target should Thibaut Courtois...

Num# 174 | Cluster #1
RT @FutbolAlAndalus: As Dani Alves leaves Barcelona, its worth reflecting that #Sevilla signed him from Bahia for £400,000. One of the grea...

Num# 175 | Cluster #1
Comienza Festival flamenco de la Mistela de Los Palacios con reconocimiento a Cristina Heeren #noticias #sevilla https://t.co/eHKa0usvWt
---
Num# 176 | Cluster #4
@piersmorgan basically a pic you'd expect a rival fan to put up yet you claim to support the club. Stop pretending you support Arsenal
---
Num# 177 | Cluster #1
#Trabajo #Sevilla Senior Information Security Analyst: PRAG-Prague, This position will focus on ... https://t.co/ZfUVV3annk @EmpleoYa_es
---
Num# 178 | Cluster #2
RT @TheEngancheRole: Turkish champions Besiktas have tabled a £7million offer for out-of-favour Liverpool defender Martin Skrtel. (Mail)

Num# 179 | Cluster #1
RT @salacustom: Nos vemos la próxima semana en #Sevilla, @quique_gonzalez& Los detectives. #TremendaBanda

Визначення рейтингів команд 25 травня (Sevillia, Liverpool, Chelsea, Arsenal)



wait to move to London, represent this special c

Num# 350 | Cluster #4
RT @Sergio_Santome: Oficial: Granit Xhaka fich
https://t.co/V9JCls0YNW

Num# 351 | Cluster #2
RT @LFC: Loris Karius has been allocated the Nc
https://t.co/GMEcaENLSe #KariusLFC https://t.c

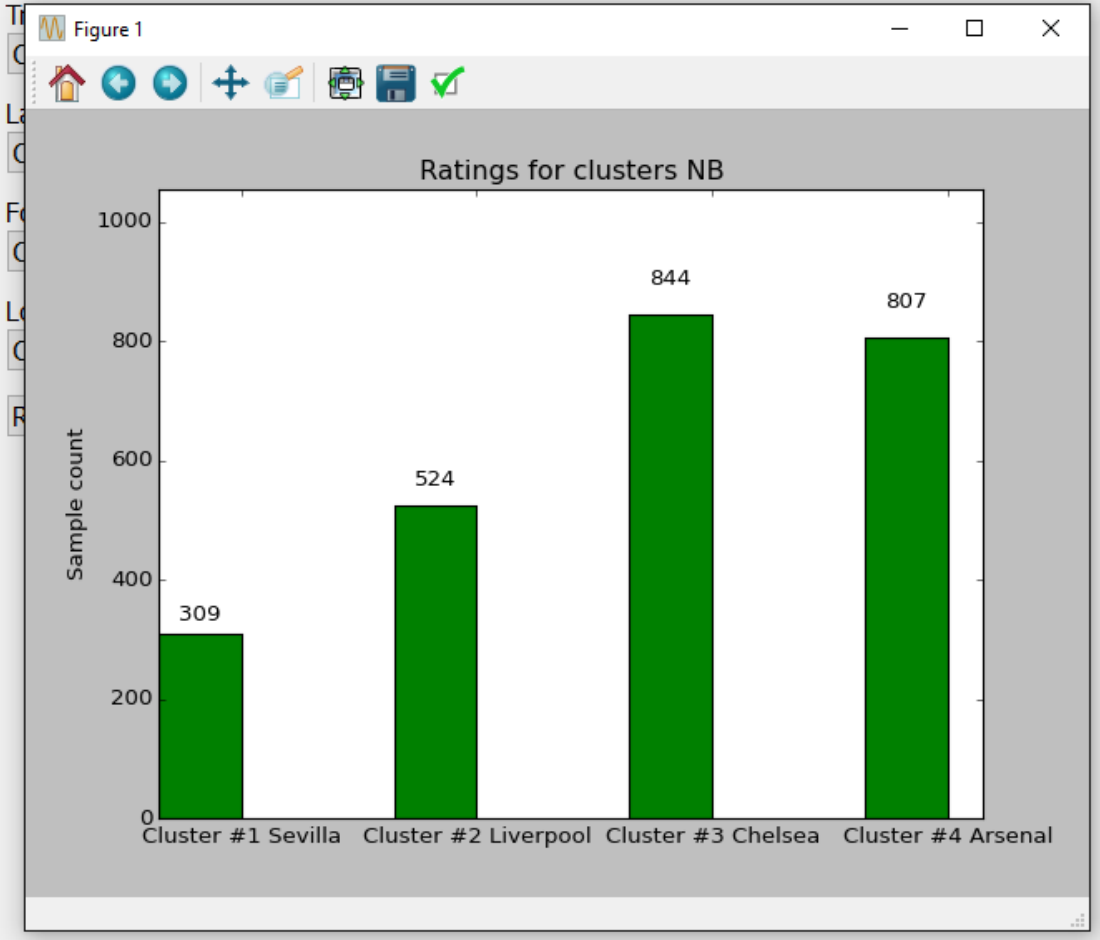
Num# 352 | Cluster #4
Welcome Xhaka!
@Arsenal how do we pronounce, it's zhaka I guess, lol
#Arsenal
#Xhaka https://t.co/fjluDO2jv9

Num# 353 | Cluster #1
@ElRinkonFamoso CONCIERTO de @_SaraMarin_ #RT
Viernes 27. Sevilla. 22h
Entradas: https://t.co/zi23WdkbYZ / taquilla https://t.co/zKJzQoBrO9

Num# 354 | Cluster #4
RT @TransferSources: BREAKING: Arsenal have completed the signing of Granit Xhaka for £30m from Monchengladbach. https://t.co/j3IKn6I8x7

Num# 355 | Cluster #4
Photo: The first official picture of Granit Xhaka in an Arsenal shirt. [@Stuart_PhotoAFC] #afc https://t.co/4wXvVPvlec

Визначення рейтингів команд 24.05-02.06 (Sevillia, Liverpool, Chelsea, Arsenal)



politicians are just as good as MPs as potential Mayors- and often bette...

Num# 12 | Cluster #3
President Obama, pardon Edward Snowden and Chelsea Manning | Trevor Timm: When it comes to civil liberties, O... https://t.co/cnRa1FtJFi

Num# 13 | Cluster #4
RT @tancredipalmeri: Arsenal have contacted Inter for Colombia international defender Murillo. Pricetag approx 20m€

Num# 14 | Cluster #4
GOOLL DE ESPAÑA!
Nolito adelanta, Taco de Pedro a Bellerin, Conducción y pase atrás del jugador del Arsenal para que Nolito defina a placer

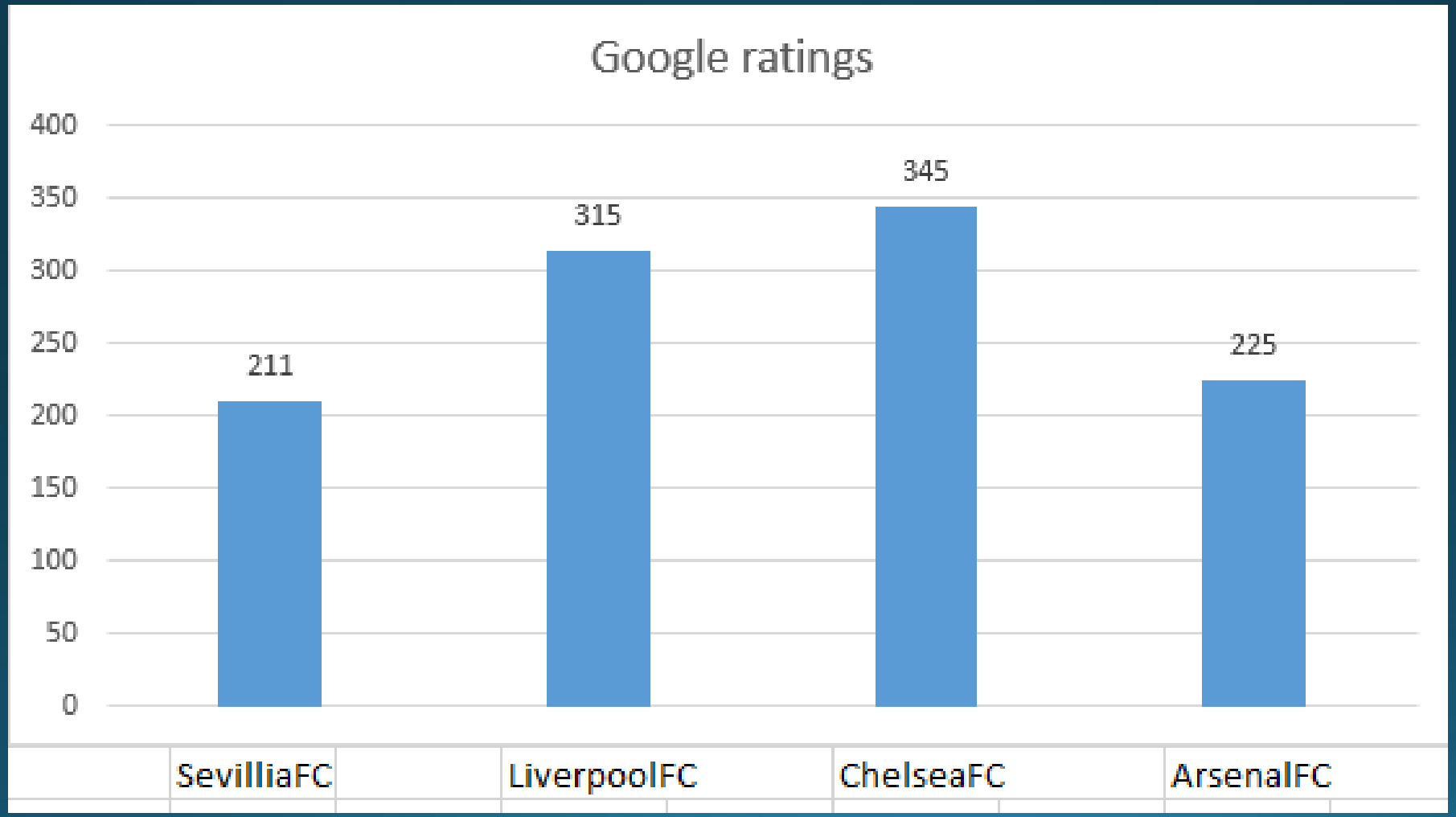
Num# 15 | Cluster #4
RT @davidhickman14: Granit Xhaka will now wear squad number 29 next season #Arsenal https://t.co/4Pf5ONIzgm

Num# 16 | Cluster #3
RT @barcastuff: Barcelona are willing to pay 40M for PSG defender Marquinhos (22). He has also offers from Man United and Chelsea, prefers...

Num# 17 | Cluster #1
RT @DeadlineDayLive: Manchester United are battling Barcelona for the signing of Sevilla midfielder Grzegorz Krychowiak. (Source: Mundo Dep...

Num# 18 | Cluster #3
@ChelseaRelated why are Chelsea players scoring the worst goals

Умовний рейтинг популярності за кількістю посилань Google



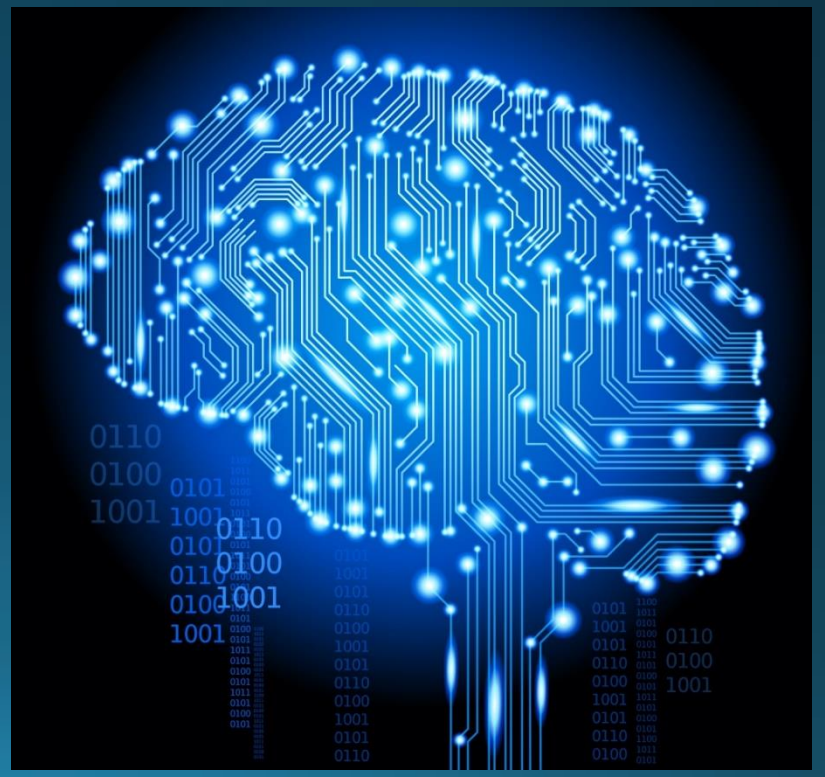
Аналіз якості отриманих результатів

К-ть класифікованих твітів\об’єм тренувальної вибірки	80	100	150
300	79%	93%	88%
500	78%	91%	87%
1000	75%	89%	85%

	Процент збереженої дисперсії в даних	Мінімальна схожість текстів в кластері	Мінімальна схожість класифікації
	30%	0,85	0,44

Наукова новизна

- В роботі вперше застосовано модель семантичної пам’яті LSA для автоматизованого формування тренувальної вибірки алгоритму машинного навчання Naïve Bayes
- Алгоритм автоматизованого формування тренувальної вибірки не залежить від предметної області і може бути використаний в різних сферах аналізу тексту



Висновки

- Розроблено автоматизовану систему аналізу соцмережі Twitter, яка здатна визначати рейтинги за популярністю у соцмережі заданих футбольних команд
- Розроблено алгоритм кластеризації текстів та автоматизованого формування тренувальної вибірки на основі моделі семантичної пам’яті LSA
- Розроблена система здатна виконувати збір даних із соціальної мережі «на льоту» та потребує мінімальних вхідних даних для розпізнавання тексту

Перспективи вдосконалення розробки

- Реалізація підсистеми для автоматизованого вибору початкових контекстів та вибору термів на основі Google Search API
- Реалізація REST сервісу на основі розробленого програмного продукту
- Робота із різними соціальними мережами та порталами даних

Дякую за увагу