МЕТОД ОТРИМАННЯ ЛЕКСИЧНОЇ ОНТОЛОГІЇ З ЕЛЕКТРОННОГО ТЕКСТУ

ВИКОНАВ: ТРАЧЕНКО С.С. 1КСУА-15 МН

КЕРІВНИК: Д.Т.Н., ПРОФ., БІСІКАЛО О. В.

ОТРИМАННЯ ЛЕКСИЧНОЇ ОНТОЛОГІЇ

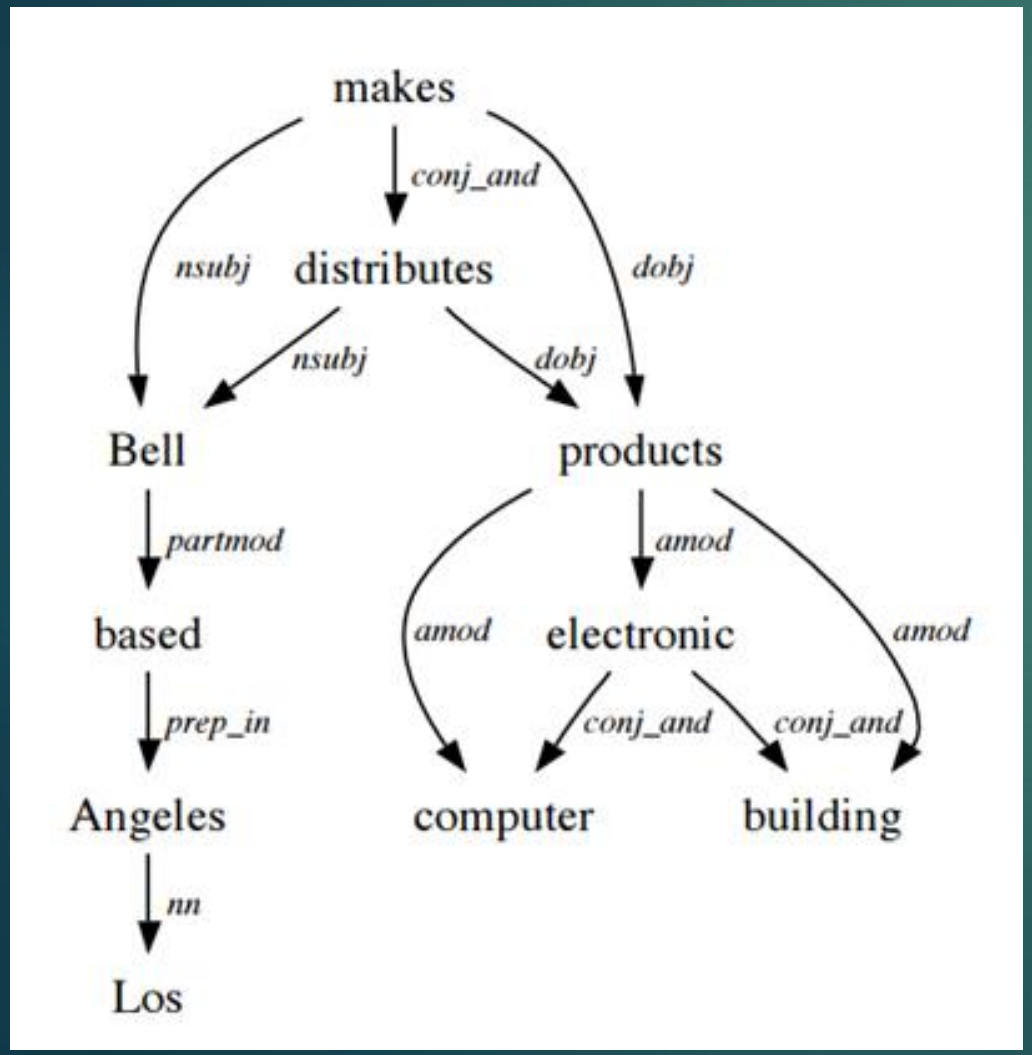
- **Актуальність дослідження**. Обсяги вільно доступної сьогодні інформації у мережі Інтернет перевищили найоптимістичніші прогнози початку нинішнього тисячоліття. І хоча в Інтернеті невпинно збільшується питома вага мультимедіа-ресурсів, природно-мовна інформація залишається найбільш важливою з точки зору реалізації глобального пошуку та рекламних функцій для користувачів. Незворотне зростання популярності лінгвістичних Інтернет-технологій вимагає у дослідників підвищення якісних показників розв'язання задач оброблення текстової інформації. Підтримку життєво важливих для розвитку мережі Інтернет функцій забезпечує цілий ряд задач комп'ютерної лінгвістики. Історично найбільш відомою з таких задач вважається визначення ключових слів тексту, але дуже близькою до неї за сутністю є задача побудови лексичної онтології цього ж тексту. Рішення даних задач дозволяє зробити висновок про загальну тему та зміст тексту, що в подальшому можна застосувати в галузі SEO – з метою визначення найбільш релевантних результатів на пошукові запити користувачів.

ОТРИМАННЯ ЛЕКСИЧНОЇ ОНТОЛОГІЇ

- **Мета роботи** полягає у підвищенні якості автоматизованої побудови лексичної онтології за рахунок визначення параметрів складних залежностей між мовними одиницями тексту.
- **Предмет дослідження** – методи та інструментальні засоби автоматизованої побудови лексичної онтології англомовного тексту.
- **Наукова новизна** роботи – уперше запропоновано метод отримання лексичної онтології з тексту, який, на відміну від відомих, базується на визначенні чисельних ознак складних зв'язків між мовними одиницями та технологічних можливостях сучасних лінгвістичних пакетів, що дозволяє гнучко будувати онтології з обраних частин мови, типів зв'язків, а також на основі різних способів обробки тексту.

ОТРИМАННЯ ЛЕКСИЧНОЇ ОНТОЛОГІЇ

Bell, based in Los Angeles, makes and distributes electronic, computer and building products.



- nsubj(makes-8, Bell-1)
- nsubj(distributes-10, Bell-1)
- vmod(Bell-1, based-3)
- nn(Angeles-6, Los-5)
- prep in(based-3, Angeles-6)
- root(ROOT-0, makes-8)
- conj and(makes-8, distributes-10)
- amod(products-16, electronic-11)
- conj_and(electronic-11, computer-13)
  amod(products-16, computer-13)
- conj_and(electronic-11, building-15)
- amod(products-16, building-15)
- dobj(makes-8, products-16)
- dobj(distributes-10, products-16)

ОТРИМАННЯ ЛЕКСИЧНОЇ ОНТОЛОГІЇ

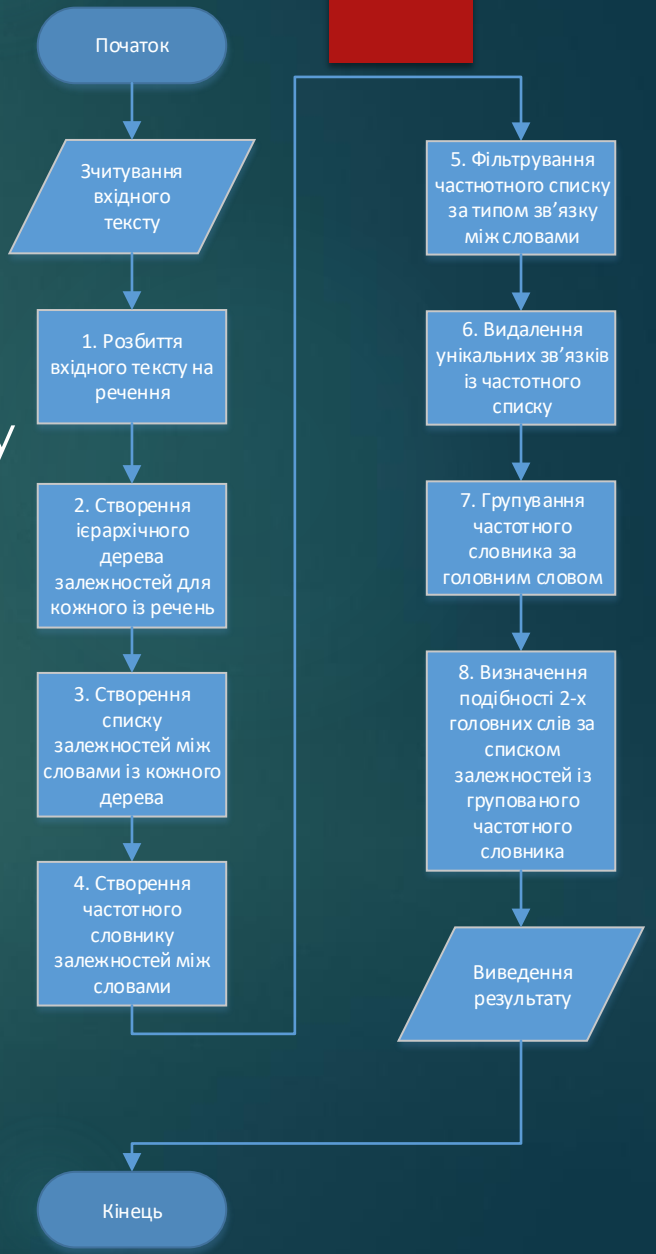
- Бібліотека NLTK, або NLTK - пакет бібліотек і програм для символьного і статистичної обробки природної мови, написаних на мові програмування Python.
- Визначення онтології в загальному складається із 2 кроків:
  1) Визначення із вхідного тексту словника залежностей, в якому ключ – головне слово, значення – список залежних слів
  2) Визначається подібність між словами (ключами словника) за частотою однакових залежних слів

Подібність двох слів(n1 та n2) S визначається формулою:

$$S = \frac{SimilarSum}{OverralSum}$$

де SimilarSum – частота однакових залежних слів між n1 та n2,

OverralSum – загальна частота залежних слів у n2 та n2



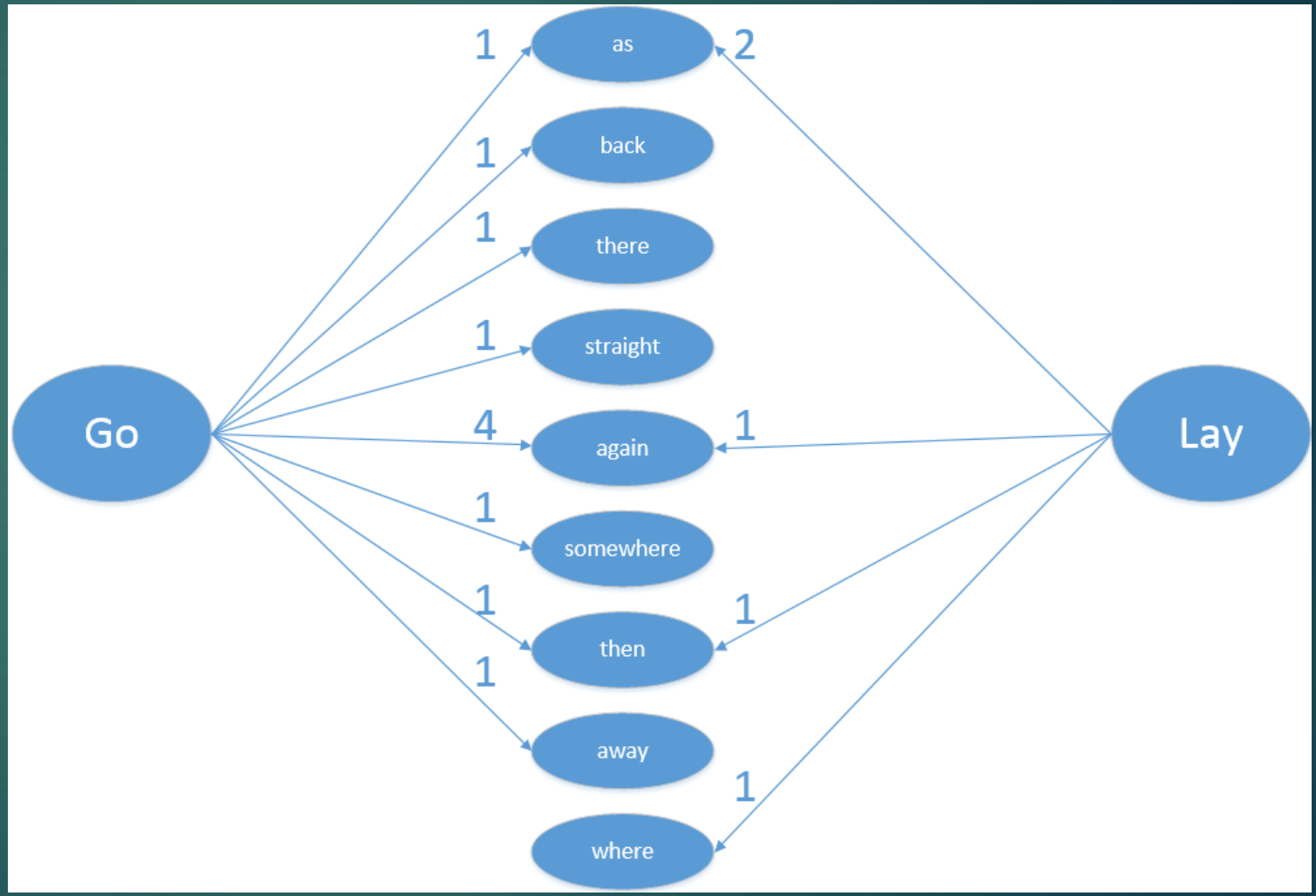
ОТРИМАННЯ ЛЕКСИЧНОЇ ОНТОЛОГІЇ

Приклад алгоритму

- OverralSum = 16
- SimilarSum = 10

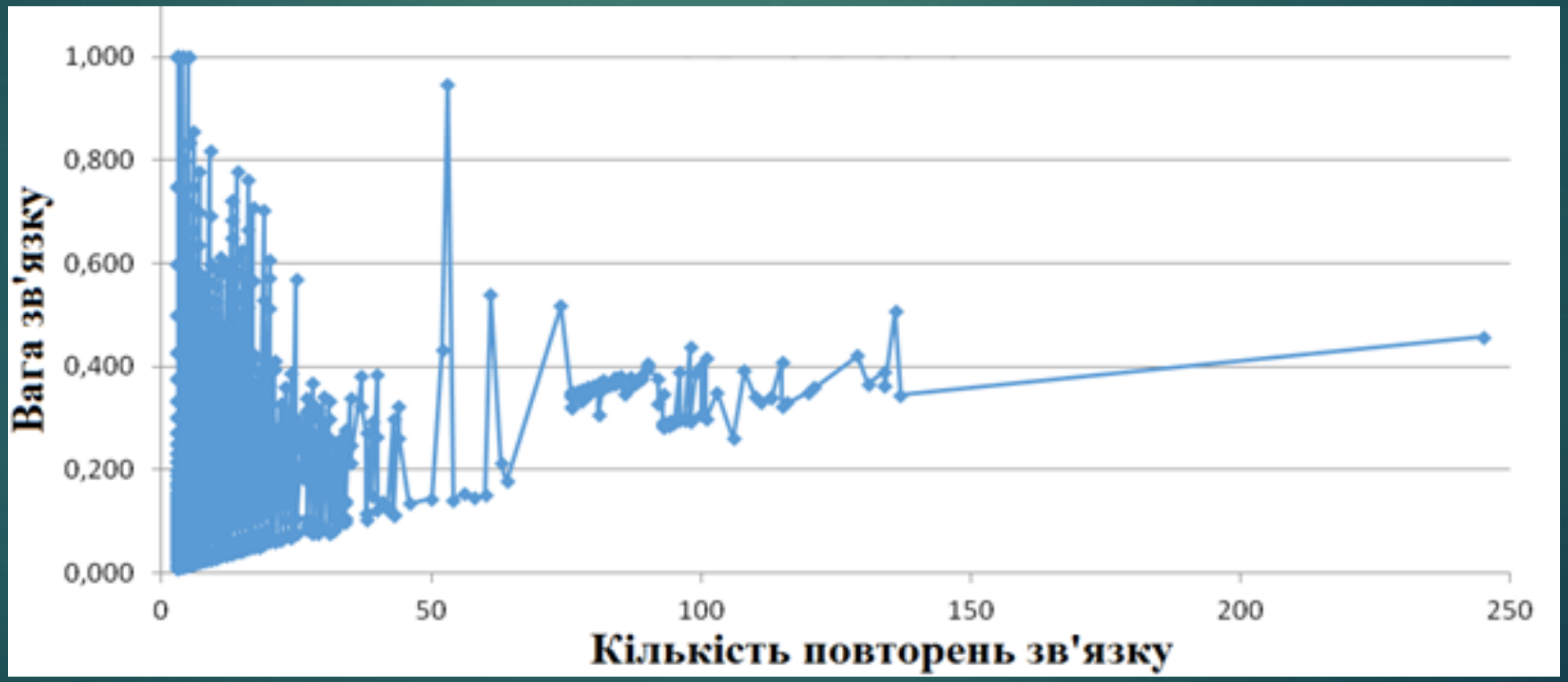
- S = 0,625

- Wordnet = 0.33



ОТРИМАННЯ ЛЕКСИЧНОЇ ОНТОЛОГІЇ

- Для експерименту було розглянуто текст роману «Мобі Дік» Германа Невіла. Отриманий розподіл частоти зв’язків має наступний вигляд:

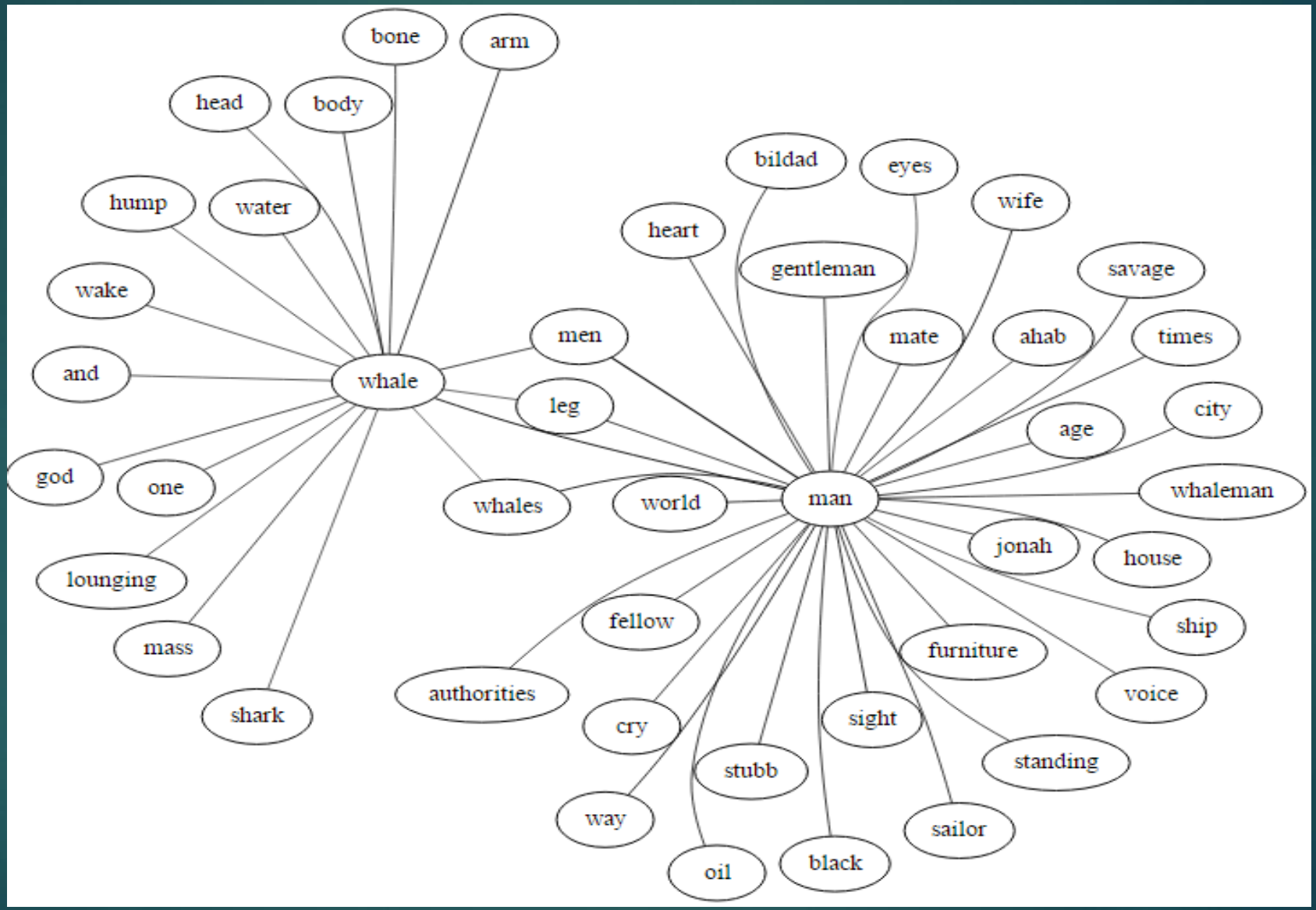


ОТРИМАННЯ ЛЕКСИЧНОЇ ОНТОЛОГІЇ

1 слово	2 слово	Кількість	Вага наша	Вага WordNet
man	whale	245	0,458	0,250
arm	whale	134	0,391	0,100
head	whale	134	0,363	0,200
men	whale	134	0,363	0,250
body	whale	131	0,367	0,125
man	whales	129	0,422	0,250
god	whale	121	0,359	0,167
mass	whale	120	0,350	0,111
water	whale	116	0,330	0,143
and	whale	115	0,323	0,000
hump	whale	113	0,342	0,077

1 слово	2 слово	Кількість	Вага наша	Вага WordNet
stub	man	92	0,377	0,000
ship	man	92	0,327	0,111
man	wife	90	0,405	0,250
man	savage	90	0,396	0,250
man	house	89	0,385	0,167
man	fellow	89	0,377	0,500
voice	man	88	0,368	0,250
man	oil	87	0,380	0,111
man	sailor	87	0,380	0,500
man	leg	87	0,366	0,143
heart	man	87	0,361	0,143

ОТРИМАННЯ ЛЕКСИЧНОЇ ОНТОЛОГІЇ



ОТРИМАННЯ ЛЕКСИЧНОЇ ОНТОЛОГІЇ

- Експеримент проводився, використовуючи наступні методи кластеризації: Spectral, SCAN, Greedy–Newman, Walktrap, LPA, Clauset–Newman, Bigclam.

- Метрики оцінювання якості графів (goodness metrics):

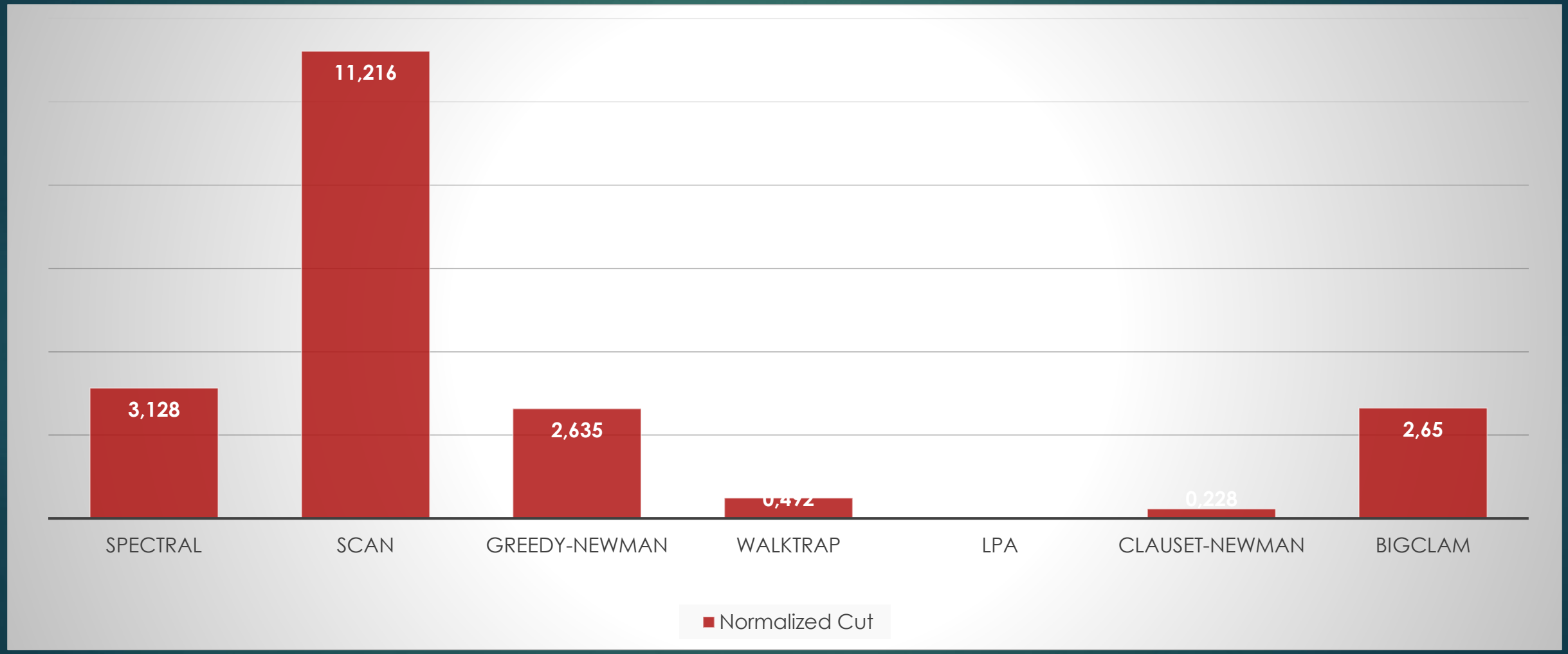
  - $\mathrm{RatioCut}\left(\hat{C}_1, \dots, \hat{C}_{N_{\hat{C}}}\right) = \sum_{i=1}^{N_{\hat{C}}} \frac{\mathrm{cut}(\hat{C}_i V \backslash \hat{C}_i)}{|\hat{C}_i|}$

  - $\mathrm{NormalizedCut}\left(\hat{C}_1, \dots, \hat{C}_{N_{\hat{C}}}\right) = \sum_{i=1}^{N_{\hat{C}}} \frac{\mathrm{cut}(\hat{C}_i V \backslash \hat{C}_i)}{\mathrm{vol}(\hat{C}_i)}$

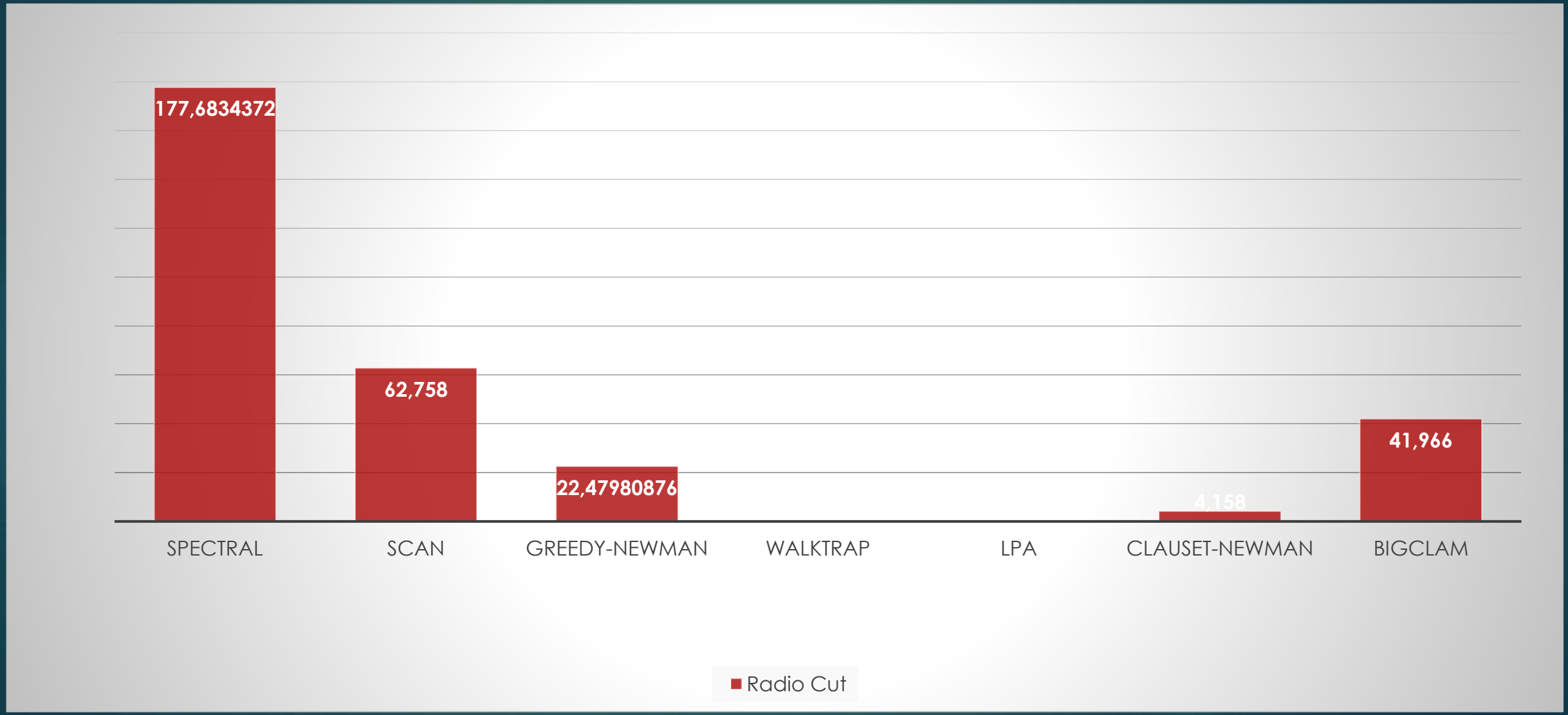
- Метрика демонструє наскільки щільні зв'язки між вершинами всередині кластера, і в той же час розріджені між кластерами (modularity)

  - $Q = \frac{1}{2m} \sum_{i,j=1}^{n} \left(\omega_{ij} - \frac{d_i d_j}{2m}\right)$

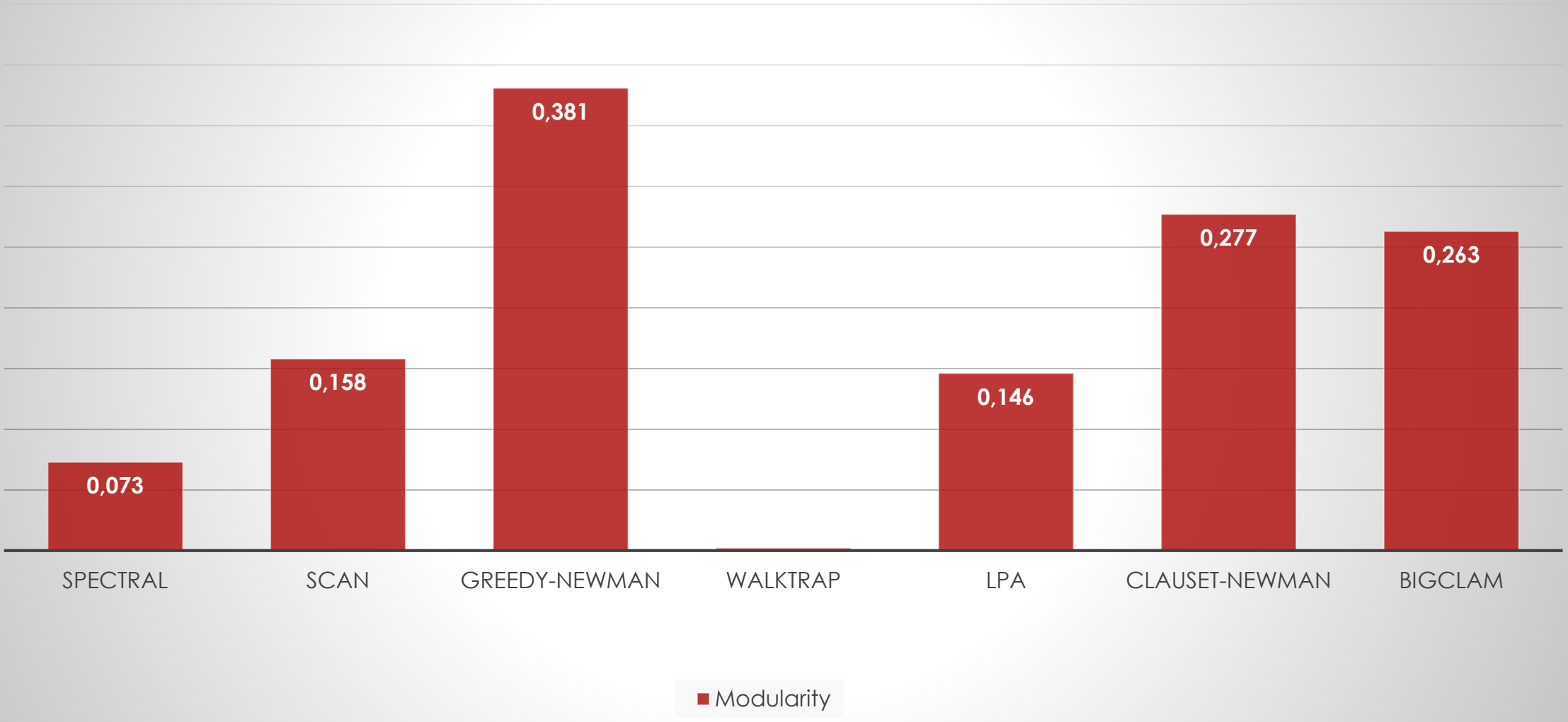
ОТРИМАННЯ ЛЕКСИЧНОЇ ОНТОЛОГІЇ



ОТРИМАННЯ ЛЕКСИЧНОЇ ОНТОЛОГІЇ



ОТРИМАННЯ ЛЕКСИЧНОЇ ОНТОЛОГІЇ



Використана література

1. About WordNet [Електронний ресурс]: – Режим доступу: http://stevenloria.com/tutorial-wordnet-textblob. – Назва з екрану
2. Метод Леска [Електронний ресурс]: – Режим доступу: https://ru.wikipedia.org/wiki/Метод_Леска. – Назва з екрану.
3. Бісікало О.В. Формалізація понять мовного образу та образного сенсу природно-мовних конструкцій / О.В. Бісікало // Математичні машини і системи. – 2012. – № 2. – С. 70–73.
4. Rank distributions of words in additive many-step Markov chains and the Zipf law [Електронний ресурс]: – Режим доступу: http://arxiv.org/pdf/physics/0406099.pdf. – Назва з екрану.
5. Каніщева О. В. Використання карт відношень (TRM) для автоматичного реферування / О. В. Каніщева // Вісник Національного університету "Львівська політехніка". – 2013. – № 770 : Інформаційні системи та мережі. – С. 108 – 122.
6. Фрумкина Р. М. Статистические методы изучения лексики / Р. М. Фрумкина. – М.:Наука, 1964. – 115 с.
7. Stanford Dependencies Manual. Revised for the Stanford Parser v. 3.5.2 in April 2015. [Електронний ресурс]. – Режим доступу: http://nlp.stanford.edu/software/dependencies_manual.pdf. – Назва з екрану
8. Herman Melville. Moby-Dick; or the Whale. [Електронний ресурс] – Режим доступу: https://www.gutenberg.org/files/2701/2701-h/2701-h.htm – Назва з екрану.

Публікації

- Бісікало О.В. Визначення змістовних ознак тексту на основі аналізу зв'язків між лексичними одиницями / О.В. Бісікало, А.І. Лісовенко, О.В. Яхимович, С.С. Траченко // Вісник національного технічного університету "ХПІ". – 2015. – № 21. – С. 83 – 85.
- Траченко С. Побудова онтології тексту природної мови за допомогою пакету NLTK / С. Траченко, О. Бісікало // Молодь в технічних науках: дослідження, проблеми, перспективи (МТН-2015) : Матеріали міжнародної Інтернет-конференції. 23–26 квітня 2015 р. – Вінниця: ТОВ «Нілан-ЛТД», 2015. – С. 66-67. – ISBN 978-966-924-027-9.
- Лісовенко А.І. Підтримка діалогу з навчальним контентом [Електронний ресурс] / А.І. Лісовенко, О.В. Бісікало, О.В. Яхимович, С.С. Траченко // Адаптивні технології управління навчанням: матеріали першої міжнародної конференції. Одеса, 23-25 вересня 2015 р. – Одеса, 2015. – С. 97-100. – Режим доступу: https://drive.google.com/file/d/0B34KZFqaGoAyOVRGTlc2clJ0eUk/view.
- Бісікало О.В. Моделювання процесів побудови парадигматичних зв'язків між словоформами на основі вимірювання текстової інформації / О.В. Бісікало, С.С.Траченко, О.В. Яхимович, А.І. Лісовенко // Вимірювання, контроль та діагностика в технічних системах (ВКДТС-2015): збірник тез доповідей III міжнар. наук. конф. (Вінниця, 27-29 жовтня 2015 р.). – Вінниця: ПП «ТД «Едельвейс і К», 2015. – С. 119-121.
- Oleg Bisikalo, Anna Lisovenko, Olexandr Jahumovuch, Sergii Trachenko, Mykola Pradivliannyi. System of Computational Linguistic on Base of the Figurative Text Comprehension / Proceedings of the 2016 IEEE First International Conference on Data Stream Mining & Processing (DSMP). – Lviv, Ukraine, August 23–27, 2016. – Publishing House of Lviv Polytechnic National University, 2016. – Pp. 69–74. – IEEE Catalog Number: CFP16J130PRT. – ISBN 978–1–5090–3735–3.

Висновки

В роботі уперше запропоновано метод отримання лексичної онтології з тексту, який, на відміну від відомих, базується на визначенні чисельних ознак складних зв’язків між мовними одиницями та технологічних можливостях сучасних лінгвістичних пакетів, що дозволяє будувати онтології з обраних частин мови, типів зв’язків, а також на основі різних способів обробки тексту.

Аналіз експериментальних даних на основі кластеризації отриманих зв'язків дозволив визначити оцінки метрик якості (goodness metrics) та метрик модулярності (modularity). Результати оцінки якості кластеризації показують, що найбільш доцільно для отримання значущих підмножин (підграфів) загальної лексичної онтології тексту застосовувати алгоритми LPA, Walktrap, Clauset–Newman та Greedy–Newman.

Дякую за увагу