

Semantrum

NLP для аналізу українських медіа

Юлія Макогон, компанія Semantrum

Компанія “Семантрум”

- входить в групу компаній “Ліга”
- Сайт: <https://promo.semantrum.net>
- Демо: <https://mediatest.semantrum.net>
- Телеграм-бот: <https://promo.semantrum.net/uk/obyrajte2019/>
- власна моніторингова онлайн-система
- веб-джерела з 1998 року
- серед клієнтів Кабінет Міністрів України, Верховна Рада України, Консультативна Місія ЄС, НАБУ, Transparency International, ДТЕК, Укравто
- українська, російська, англійська мови, а також деякі мови ЄС



1. МОНІТОРИНГ ДЖЕРЕЛ

Semantrum - це онлайн-система, у якій у режимі 24/7 накопичуються повідомлення з таких джерел:

20 загальноукраїнських **телеканалів**

500+ друкованих **ЗМІ** (загальноукраїнські та регіональні)

10 000+ **сайтів** новин, інформаційних агентств, онлайн-представництв держорганів та компаній (українська, російська, англійська та інші мови)

15 українських **радіостанцій**

Соціальні мережі: Facebook, Twitter, YouTube, LiveJournal, vKontakte, Odnoklassniki

Сайти з **відгуками**

Сайти з **резюме**





2. АВТОМАТИЗОВАНІЙ КОНТЕНТ-АНАЛІЗ

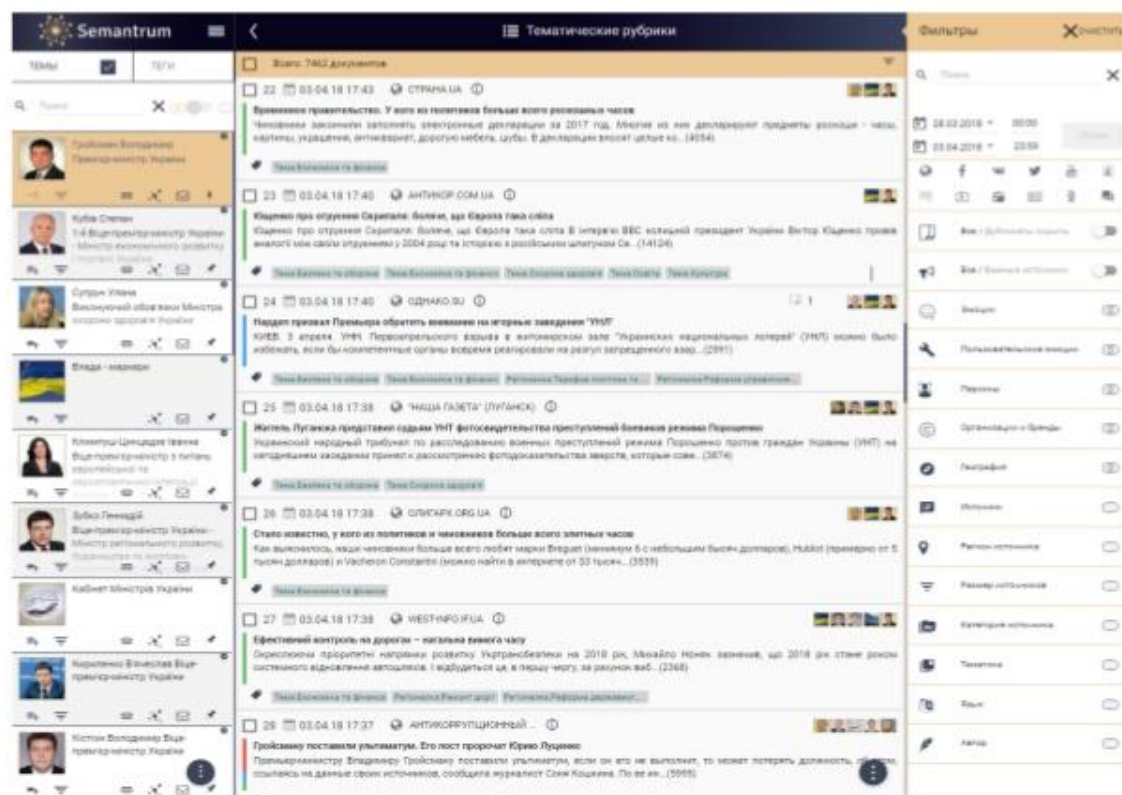
Формування стрічок та накопичення повідомлень за Вашими **темами**

Визначення **тональності** згадування Вашого об'єкту моніторингу

Автоматична **класифікація** кожного повідомлення за Вашими темами

Виділення у повідомленнях будь-яких **персон, брендів, компаній, географічних назв**

Визначення **повідомлень-дублів** та схожих публікацій



Робота з окремим повідомленням

В інтерв'ю **Радіо Свобода** **Анатолій Гриценко** піддав критиці дії **Володимира Гройсмана**

Анатолій Гриценко, лідер партії «Громадянська позиція», розповів, що в нього була професійна розмова в студії з **Гройсманом**.

«Я дорікав прем'єр-міністрові, хотів, щоб люди це знали. Щоб люди знали, що не можна показувати типу економічні здобутки уряду, набираючи майже під 10 відсотків кредити на 10 років в твердій валюті», – почав **Гриценко**.

«Тобто **Гройсман** піде, а всі оці всі борги залишаться, розумієте?», – каже **Анатолій Гриценко**.

- визначення мови
- токенизація
- класифікація
- пошук дублікатів статті та статей з тим же сюжетом
- виділення іменованих сутностей (NER)
- зв'язування іменованих сутностей (Named Entity Linking, NEL)
- визначення тональності повідомлення і тональності згадки про сутність

Ресурси для роботи з українською

- lang_uk: <http://lang.org.ua/uk/>
- колекція дерев залежностей Universal Dependencies (UD) <https://universaldependencies.org/>
- spaCy <https://spacy.io/>

Для англійської всього більше...

Dear God, please give
me more money or
take away my
expensive taste.



somee cards
user card

Піраміда Маслоу NLP

парсер залежностей

частини мови POS, визначення
іменованих сутностей NER

лематизація, вектори


визначення мови, токенизація



Порівняння UD за кількістю токенів

German	3409000		Dutch	307000	Greek	63000
Czech	2222000		Latvian	208000	Old Church Slavonic	57000
Japanese	1688000		Croatian	199000	Classical Chinese	55000
Russian	1263000		Swedish	195000	Gothic	55000
French	1156000		Old French	170000	Afrikaans	49000
Arabic	1042000		Slovenian	170000	Maltese	44000
Spanish	1004000		Galician	164000	Wolof	44000
Italian	781000		Old Russian	164000	Vietnamese	43000
Norwegian	666000		Chinese	161000	Hungarian	42000
English	603000		Hebrew	161000	Lithuanian	42000
Latin	582000		Bulgarian	156000	Uyghur	40000
Portuguese	570000		Persian	152000	Armenian	36000
Catalan	531000		Indonesian	141000	Hindi English	26000
Polish	500000		Urdu	138000	North Sami	26000
Estonian	461000		Ukrainian	122000	Coptic	25000
Romanian	460000		Basque	121000	Irish	23000
Korean	446000		Slovak	106000	Thai	22000
Ancient Greek	416000		Danish	100000	Erzya	15000
Finnish	377000		Serbian	97000	Bambara	13000
Hindi	375000		Turkish	91000	Belarusian	13000

UD за кількістю речень

		sentences	tokens	tokens / sentences
	German	191757	3409000	17,78
	Russian	71183	1263000	17,74
	English	34631	603000	17,41
	Polish	40454	500000	12,36
	Dutch	20924	307000	14,67
	Ukrainian	7060	122000	17,28
	Turkish	9437	91000	9,64
	Vietnamese	3000	43000	14,33
	Hungarian	1800	42000	23,33

“Нормальна” українська

- зміни у правописі



24 КАНАЛ

НОВИЙ ПРАВОПИС: ЯК ПРАВИЛЬНО

- И НА ПОЧАТКУ СЛОВА**
✓ індик та индик ✓ ирій та вирій
- БІЛЬШЕ Ѓ У СЛОВАХ**
- БІЛЬШЕ ЕТЕРІВ**
✓ анафема та анатема ✓ міф та міт
✓ ефір та етер ✓ Афіни та Атени
- ПОВЕРНЕННЯ ЙОТУВАННЯ**
[j] із голосними буде передаватись
буквами є, ї, ю, я: ✓ проєкт ✗ проект
- МЕНШЕ ДЕФІСІВ**
✓ попмузика ✓ вебсторінка ✓ пресконференція
✗ поп-музика ✗ веб-сторінка ✗ прес-конференція
- ПІВ – ОКРЕМО**
✓ пів Києва ✓ пів яблука ✓ пів години
✗ пів-Києва ✗ пів' яблука ✗ півгодини
Але: півострів, півзахист або півоберт
- ДИФТОНГИ АУ – АВ**
✓ аудиторія та авдиторія
✓ пауза та павза
- ОНОВЛЕНІ ВЕЛИКІ БУКВИ**
Релігійні поняття – пишемо з великої:
✓ Бог ✓ Трійця
✗ бог ✗ трійця
Посади українських держслужбовців – з маленької:
✓ президент ✗ Президент
- ФЕМІНІТИВИ:**
Використовуємо: директорка, учениця, філологиня, поетеса

“Нормальна” українська

- зміни у правописі
- синтетичність

В Україні буде
«ТИПОВА
груднева
какабека» –
Діденко

“Нормальна” українська

- зміни у правописі
- синтетичність
- суржик
- орфографічні помилки
- Google Translate

В Україні 355 політичних партій. Нездається вам, що це занадто багато. Ми вже обішли в цьому весь цивілізований світ, а рівень життя все зменшується, а кількість патрій збільшується. Тут є питання; Україна, що полігон для випробування політтехнологій? Досить цих експериментів потрібно наводити лад в державі! партія повинна мати чітку ідеологію, а в нас партії Кличка, Ляшка, Тимошенко, Бойка та інших людей та залежно від їх характеру визначається ідеологія діяльності . Так неповинно бути!!! Має бути прописана чітка ідеологічна стратегія ті чи іншої партії та шляхи її досягнення.

Токенізація

- tokenize_uk <https://github.com/lang-uk/tokenize-uk>
- spaCy <https://spacy.io/>
 - без моделі: Sentencizer
 - натренувати парсер залежностей на основі https://github.com/UniversalDependencies/UD_Ukrainian-IU
<https://spacy.io/api/cli#convert> --n-sents, -n

Представлення слів

- стемер https://github.com/Amice13/ukr_stemmer
- лематизатор <https://github.com/kmike/pymorphy2> (ставити треба не через pip, а з github!)

```
pip install git+https://github.com/kmike/pymorphy2.git  
pymorphy2-dicts-uk
```

- вектори Word2Vec, Glove, Lex2Vec <http://lang.org.ua/en/models/#anchor4>
- fasttext
- Bpemb <https://github.com/bheinzerling/bpemb>
- BERT <https://github.com/google-research/bert/blob/master/multilingual.md>

spaCy

- <https://spacy.io/>, є моделі лінгвістичних анотацій для 9 європейських мов, а також компоненти для класифікації (TextCategorizer) і розбору документу за правилами (Matcher, EntityRuler і т.д.)
- починаючи з версії 2.1, підтримує токенізацію та лематизацію та створення моделей для української мови
- лематизація для української та російської базується на <https://github.com/kmike/pymorphy2>
- лематизація працює коректно, якщо є натренований POS-tagger; без моделі лишає форму слова якщо багато варіантів
- натренувати модель з частинами мови та парсером залежностей можна на UD
- файли UD, готові до тренування https://github.com/juliamakogon/spacy_ud_uk

Результати для моделі, з векторами lang_uk

На UD:

POS 96.44

UAS 84.02

LAS 79.30

На наших даних:

NER P 86.67

NER R 87.46

NER F 87.06

Як тренувати модель для spaCy

<https://spacy.io/usage/training>

- у коді
- через командний рядок <https://spacy.io/api/cli>

```
python -m spacy convert [input_file] [output_dir] [--file-type]  
[--converter] [--n-sents] [--morphology] [--lang]
```

```
python -m spacy init-model [lang] [output_dir] [--jsonl-loc]  
[--vectors-loc] [--prune-vectors]
```

```
python -m spacy train [lang] [output_path] [train_path] [dev_path]  
[--base-model] [--pipeline] [--vectors] [--n-iter]  
[--n-early-stopping] [--n-examples] [--use-gpu]
```

--pipeline, -p Comma-separated names of pipeline components to train. Defaults to 'tagger,parser,ner'.

```
python -m spacy evaluate [model] [data_path] [--displacy-path]  
[--displacy-limit]  
[--gpu-id] [--gold-preproc] [--return-scores]
```

Тренуємо статистичну модель для POS і лематизації

```
pip install -U spacy
```

```
pip install git+https://github.com/kmike/pymorphy2.git  
pymorphy2-dicts-uk
```

```
python -m spacy train uk ttt ud_uk/uk_iu-ud-train.json  
ud_uk/uk_iu-ud-dev.json -p tagger
```

```
python -m spacy evaluate ttt/model-final ud_uk/uk_iu-ud-test.json
```

Як працювати зі spaCy та українською

```
import spacy
```

```
# без моделі
```

```
nlp = spacy.blank('uk')
```

```
sentencizer = nlp.create_pipe("sentencizer")
```

```
nlp.add_pipe(sentencizer)
```

```
# завантажити модель
```

```
nlp = spacy.load(PATH)
```


Заважає відсутність даних для NER.

Що можна зробити:

- проанотувати свої дані - витрати часу та фінансів
- ner_uk - мало даних, можна сумістити зі [SpaCy EntityRuler](#)
- transfer learning: на основі Multilingual BERT [DeepPavlov](#)
- автоматична анотація на основі Вікіпедії подібно до WikiNER

Розмічені дані необхідні для навчання та валідації



<https://twitter.com/richardsocher/status/840333380130553856>

Підсумки

- Українська мова цікавить замовників: медіа-аналітика, рітейл і т.д.
- Інструменти для української мови є і з'являються нові
- Необхідно більше даних для валідації алгоритмів та навчання
- Компанія “Семантрум” рада співробітництву для розвитку засобів аналізу української мови