Raport techniczny

Eksplorator Informacji – nie młotek, a skalpel

Wersja 1.0

Team: radlab.dev

14.07.2025

Streszczenie

Dokument zawiera techniczny opis funkcjonalności *Przeglądarka Informacji* oraz *Eksplrator Informacji* znajdujących się na naszym *playgroundzie*. Przedstawione zostały dokładne definicje poszczególnych elementów, wskazane zostały zależności, struktury danych, ustawienia algorytmów i wykorzystane metody. Ten dokument ma być pewnego rodzaju podsumowaniem prac i przygotowaniem *playgroundu* do dalszego rozwoju o nowe metody analizy i wizualizacji danych. Kolejne sekcje opisują coraz dokładniej poruszany problem i rozwiązanie, poczynając od ogólnej definicji informacji, definicji informacji dziennej, ciągłej oraz nieciągłej. Ostatni rozdział, to swojego rodzaju podsumowanie planów na dalsze prace.

1 Intro

Od niedawna na naszym *playgroundzie* działają dwa rozwiązania Przeglądarka Informacji oraz Eksplorator informacji. To narzędzia, które pomagają spojrzeć na pojawiające się newsy z szerszej perspektywy – pozwalają analizować jak dana wiadomość wpasowuje się w ogólny trend informacyjny. Lub z drugiej strony: jak pojawiające się newsy tworzą bańki informacyjne. Działanie Przeglądarki i Kreatora nastawione jest na analizę informacji, nie pojedynczych wiadomości, a wiadomość wpasowana jest w informację. Przykładowe ogólne informacje: *Wojna na Ukrainie*, *Pogoda*, *Wybory prezydenckie w Polsce*. Można powiedzieć, że informacja w pewnym stopniu definiuje tematykę, która z nią jest powiązana. Jednak z innej strony, informacja może być lokalna np. *Druga tura wyborów prezydenckich*, która pojawia się w określonym czasie i kończy w momencie, kiedy przestanie się o niej pisać. Do analizy *lokalnych* i *globalnych* informacji, służą właśnie wspomniane rozwiązania, a jako *źródło zasilania* metod danymi, we wspomnianych rozwiązaniach, wykorzystany został *Strumień Aktualności* z *playgroundu*, w którym analizujemy informacje pojawiające się na ponad 20 polskich i zagranicznych stronach newsowych.

2 Login

Podstawą działania Przeglądarki i Eksploratora informacji są newsy pojawiające się codziennie w mediach. Zdefiniujmy zatem pojedynczy, dowolny news jako n w dniu d za pomocą: $n(d) = \text{news}(d)$ zaś konkretny news w dniu oznaczmy za pomocą $n_i(d) = news_i(d)$, wszystkie newsy w danym dniu oznaczmy za pomocą $N(d)$, zaś liczba dni, to z:

$$\begin{aligned} n_i(d) &= \text{news}_i(d),\ i \in \langle 0, k\rangle, k = |N(d)| \\ N(d) &= \{n_1(d), ..., n_k(d)\},\ d \in \langle d_1, ..., d_z\rangle \end{aligned} \tag{1}$$

a konkretny news w konkretnym dniu oznaczymy za pomocą $n_i(d_j)$, gdzie i wskazuje na newsa, a j na dzień. W ten sposób można zdefiniować zbiór danych, który będzie analizowany w dalszej części. Czyli newsy, które pojawiają się codziennie na różnych stronach internetowych (newsowych). Z naszej perspektywy informacja rozsiana jest w wielu newsach, zatem informacja I integruje w jednym miejscu różne n_i i co najważniejsze I zawsze posiada datę początkową i końcową, w odróżnieniu od daty $n_i(d)$ – news pojawia się w danym dniu, zaś informacja rozsiana jest po wielu dniach. Wyróżniamy dwa rodzje informacji (jednak w dalszej części, dla prostoty definicji, wprowadzimy pomocniczy typ informacji): ciągłą – CI (lokalną) i nieciągłą – NCI (globalną). Zarówno CI jak i NCI zawierają dzień rozpoczęcia d_b oraz dzień zakończenia d_e. Czyli:

$$\begin{aligned} CI_{d_b}^{d_e} &\subset \{n_i(d_b),\ n_i(d_{b+1}),\ n_i(d_{b+2}),\ ...,\ n_i(d_e)\},\ e \geq b \\ NCI_{d_b}^{d_e} &\subset \left\{CI_{d_b}^{d_{e'}}, ..., CI_{d_{b'}}^{d_e}\right\},\ e \geq b' \geq b,\ e \geq e' \geq b \\ |CI| &\leq |N(d)|,\ |NCI| \leq |CI| \end{aligned} \tag{2}$$

Gdzie d_b (b - begin) oznacza datę początkową, a d_e (e – end) datę końcową informacji I. W ten sposób zdefiniowaliśmy czym jest news, czym jest informacja oraz jak je rozróżniamy.

3 Kombo

3.1 Główne definicje

Przejdźmy teraz z opisu ogólnego, do bardziej konkretnego. Zacznijmy od przedefiniowania $n_i(d)$ do postaci wektorowej. Wcześniejszy news, aktualnie reprezentowany jest przez wektor $\vec{n}_i(d)$, który tworzony jest poprzez konwersję newsa przez model bi-enodera $\mathcal{E}$ do wektora $\vec{n}$ o wielkości 1024. Dodatkowo wprowadzony jest proces redukcji $\vec{n}$ do $\vec{r}$ metodą $\mathcal{T}$ do rozmiaru $|\vec{r}| \lll |\vec{n}|$ (definiowany jest z-ety dzień):

$$\begin{aligned} \mathcal{E}(n_i(d)) &\Rightarrow \vec{n}_i(d),\ |\vec{n}_i(d)| = 1024 \\ \vec{\mathcal{N}}(d_z) &= \{\vec{n}_1(d_z), .., \vec{n}_k(d_z)\}\ \ |\vec{\mathcal{N}}(d_z)| = |N(d_z)| \\ \mathcal{T}(\vec{n}(d)) &\xrightarrow{\text{t-sne}} \vec{r}(d),\ |\vec{r}(d)| = 2 \\ \mathbb{R}(d_z) &= \{\vec{r_1}(d_z), ..., \vec{r_k}(d_z)\},\ |\mathbb{R}(d_z)| = |\vec{\mathcal{N}}(d_z)| \end{aligned} \tag{3}$$

Gdzie $\mathcal{T}$ to t-sne z redukcją do 2 komponentów z perplexity 30, kalkulacja gradientu za pomocą aproksymacji barnes hut oraz oceną za pomocą podobieństwa cosine. Dzięki redukcji $\vec{n}_i(d)$ do 2 wymiarowego wektora $\vec{r}_i(d)$ czas potrzebny na ekstrakcję I jest zdecydowanie krótszy niż w przypadku przetwarznia 1024 wymiarowych embeddingów, a to umożliwia codzienne przeliczanie macierzy informacji $\mathbb{R}$ i ekstrakcję CI. Przejdźmy teraz do definicji Informacji w przestrzeni wektorowej, niech $\vec{I}$ oznacza wektor dowolnej informacji, odpowiednio ciągłej $\vec{CI}_{d_b}^{d_e}$ oraz nieciągłej $N\vec{C}I_{d_b}^{d_e}$, dla których zachowane są te same właściwości, jak w definicji (2) informacji:

$$\begin{aligned} \vec{CI}_{d_b\, w\times 2}^{d_e} &\subset \{\vec{r}_i(d_b), \vec{r}_i(d_{b+1}), \vec{r}_i(d_{b+2}), ..., \vec{r}_i(d_e)\},\ e \geq b,\ [c \times 2] \\ N\vec{C}I_{d_b\, h\times w\times 2}^{d_e} &\subset \left\{\vec{CI}_{d_b}^{d_{e'}}, ..., \vec{CI}_{d_{b'}}^{d_e}\right\},\ e \geq b' \geq b, e \geq e' \geq b \\ &|\vec{CI}| \leq |N(d)|,\ |N\vec{C}I| \leq |\vec{CI}| \end{aligned} \tag{4}$$

gdzie w definiuje wielkość informacji (liczba wierszy w macierzy – liczba $n_i(d_j)$ w informacji), 2 to wymiar $\vec{r}$, zaś h to liczba informacji ciągłych składających się na informację nieciągłą. Dla uproszczenia, przyjmijmy macierzowy zapis informacji:

$$\begin{aligned} \mathbb{CI} &= \vec{CI}_{d_b\, w\times 2}^{d_e},\ |\mathbb{CI}| = [w \times 2] \\ \mathbb{NCI} &= N\vec{C}I_{d_b\, h\times w\times 2}^{d_e},\ |\mathbb{NCI}| = [h \times w \times 2] \end{aligned} \tag{5}$$

Teraz najważniejsze: jak powstaje $\mathbb{CI}$ oraz $\mathbb{NCI}$? Zacznijmy od procesu powstawania $\mathbb{CI}$, który jest kilkuetapowy.

3.2 Informacja dzienna $\mathbb{DI}$

Dla lepszego zdefiniowania $\mathbb{CI}$ musimy wprowadzić nowy typ informacji, nazwijmy ją informacją dzienną, którą oznaczmy przez DI, a jej postać macierzową za pomocą $\mathbb{DI}$. Co istotne, na DI składają się takie $n_i(d)$, gdzie d dla dowolnych $n_i(d)$ oraz $n_j(d)$ jest takie same – innymi słowy, pochodzą z tego samego dnia, gdzie dla $\mathbb{CI}$ oraz $\mathbb{NCI}$ ten warunek nie musi być spełniony (rzadko kiedy jest spełniony, a w szczególności dla $\mathbb{NCI}$). W takim razie jak powstaje konkretny $\mathbb{DI}(d_i)$ wykorzystując wcześniej zdefiniowane $\mathbb{R}(d)$:

$$\begin{aligned} \mathbb{H}(d_i) = \mathcal{H}(\mathbb{R}(d_i)) &\xrightarrow{hdbscan} \vec{l}(d_i),\ |l(d_i)| = |\mathbb{R}(d_i)| \\ \forall \vec{r}_k \in \mathbb{R}(d_i) : l_k &= \text{label}(\vec{r}_k) \\ k &\in \langle -1, |\mathbb{R}(d_i)|) \\ l_k &= \begin{cases} -1 & \text{gdy outlier} \\ \geq 0 & \text{gdy inaczej} \end{cases} \end{aligned} \tag{6}$$

W równaniu (6) zdefiniowaliśmy funkcję $\mathcal{H}$ mapującą elementy $\mathbb{R}(d_i)$ na przestrzeń labelek $\vec{l}$. Do tego celu wykorzystaliśmy w podstawie metodę hdbscan dzięki której otrzymujemy przypisane identyfikatory do $\vec{r}_k$, a pośrednio do

$n(d_k)$. HDBSCan to algorytm nienadzorowanego podziału danych, który nie wymaga definiowania docelowej liczby klas, a wymaga podania hiperparametrów do tworzenia modelu $\mathbb{H}(d_i)$. Po wyuczeniu staje się funkcją mapującą dowolny $\vec{r}$ na identyfikator l_i. Jeżeli dwa $n_i(d)$ oraz $n_j(d)$ posiadają taki sam identyfikator l_k, to znaczy, że należą do tej samej grupy. Dodatkowo, wprowadziliśmy proces optymalizacji wyboru podziałów danych z różnie dobranymi hiperparametrami do tworzenia $\mathbb{H}(d_i)$:

$$\begin{aligned} \mathbb{H}(d) =& \{\mathbb{H}_1(d_i), ..., \mathbb{H}_l(d_i)\} \\ \operatorname{argmax} f(\mathbb{H}(d)) =& \{\mathbb{H}(d) : \max f(\mathbb{H}_i(d)\} \end{aligned} \tag{7}$$

Wybieramy ten podział, który maksymalizuje funkcję oceny $f(\mathbb{H}(d))$, czyli uwzględnia górną, dolną oraz optymalną liczbę grup, a przy tym redukuje podziały z największą liczbą odrzuconych przykładów – outlierów. Zatem, $\mathbb{H}(d_i)$ w definicji (6) jest de facto najlepszym modelem podziału $N(d_i)$. Elementy posiadające ten sam identyfikator l_k przynależą do tej samej grupy $\mathbb{C}_k(d)$:

$$\begin{aligned} \mathbb{C}_i(d) &= \{n_1(d), ..., n_i(d)\}\,,\ \text{label}(\vec{r}_1(d)) \equiv \text{label}(\vec{r}_j(d)) \\ \mathbb{C}(d) &= \{\mathbb{C}_1(d), ..., \mathbb{C}_h(d)\}\,,\ h = |\text{unique}(\text{label}(\mathbb{R}(d)))| \end{aligned} \tag{8}$$

czyli $\mathbb{C}_i(d)$ to zbiór newsów $n(d)$, które posiadają wspólne cechy podobieństwa i w procesie podziału zostały przypisane do tej samej grupy. Liczba informacji $|C(d)|$ jest dynamiczna, uzależniona od liczby pojawiających się newsów, jednak zawsze optymalizowana jest w kierunku dość szczegółowych powiązań semantycznych. W tej chwili $\mathbb{C}(d)$ wskazuje na dzienny podział newsów na grupy, jednak do pełnej definicji $\mathbb{DI}$ potrzebujemy jeszcze opisowo przedstawić czym jest $\mathbb{C}_i(d)$.

$$\begin{aligned} \mathcal{RC}_i(d) &= \text{rand}\, f(x : \mathbb{C}_i(d)),\ |\mathcal{RC}_i(d)| = h,\ h = 15 \\ \mathcal{T}_i(d) &= \mathcal{G}(\mathcal{RC}_{\text{i}}(\text{d})) \end{aligned} \tag{9}$$

Do tego celu, dla każdej $\mathbb{C}_i(d)$ losowo wybieramy próbę $h = 15$ elementów (newsów ze strumienia aktualności) $\mathcal{RC}_i(d)$. Następnie, wykorzystując model $\mathcal{G}$ `google-gemma3-12b-it`, na ich podstawie powstaje opisowy tekst $\mathcal{T}_i(d : \text{text})$ dla informacji $\mathbb{C}_i(d))$. Tym sposobem mamy komplet potrzebnych zależności do zdefiniowania i-tej $\mathbb{DI}$ w dniu d z dostępnych k informacji w tym dniu:

$$\begin{aligned} \mathcal{DI}_i(d) &= (\mathcal{T}_i(d), \mathcal{RC}_i(d), N_i(d)) \\ \mathbb{DI}_i(d) &= \left(\mathcal{E}(\mathcal{T}_i(d)), \vec{\mathcal{RC}}_i(d), \vec{N}_i(d)\right) \\ \mathbb{DI}(d) &= \{\mathbb{DI}_1(d), ..., \mathbb{DI}_k(d)\} \end{aligned} \tag{10}$$

I to właśnie $\mathbb{DI}_i(d)$ jest konkretną informacją wyświetlaną w Przeglądarce informacji. Poszczególne elementy trójki z definicji (10) to:

- $\mathcal{T}_i(d)$ – to tekstowe podsumowanie informacji w wybranym dniu;
- $\mathcal{RC}_i(d)$ – to próba newsów (odnośników do oryginalnych artykułów), na podstawie których postało $\mathcal{T}_i(d)$;

- $N_i(d)$ to wykaz wszystkich newsów z danego dnia, które wskazują na informację z tekstową postacią $\mathcal{T}_i(d)$;

Zaś $\mathbb{DI}(d)$ to macierzowa postać informacji, kolejno są to:

- $\mathcal{E}(\mathcal{T}_i(d))$ – embedding tekstowej postaci $\mathcal{T}_i(d)$;
- $\vec{\mathcal{RC}}_i(d)$ – macierz zredukowanych emebdingów artykułów ze Strumienia Aktualności pochodzących z próby $\mathcal{RC}_i(d)$;
- $\vec{N}_i(d)$ – to (niezredukowana) macierz reprezentująca wszystkie newsy $N_i(d)$, które opisane zostały informacją $\mathcal{T}_i(d)$;

3.3 Informacja ciągła $\mathbb{CI}$

Przejdźmy zatem do definicji informacji ciągłej $\mathbb{CI}$, dla której podstawą jest informacja dzienna $\mathbb{DI}(d)$. Można powiedzieć, że informacją ciągła, to taka informacja dzienna, która trwa dłużej niż jeden dzień (lub w skrajnym przypadku właśnie jeden dzień). Ważne jednak, że ta informacja poruszana jest dnia poprzedniego i/lub następnego w stosunku do dnia badania tej informacji. Wszystkie zależności formalne, przedstawione zostały w definicji (2) jako $CI_{d_b}^{d_e}$. Postać wektorowa $\vec{CI}_{d_b\, w\times 2}^{\,d_e}$ w definicji (4) oraz uogólnia postać $\mathbb{CI}$ na (5), dla których d_b oznacza datę rozpoczęcia informacji, zaś d_e to data zakończenia informacji, dla których zachodzi zależność $e \geq b$, czyli dzień zakończenia musi być większy (lub taki sam – jako szczególny przypadek $\mathbb{CI}$) od dnia rozpoczęcia propagacji informacji w serwisach newsowych.

3.3.1 Podobieństwo informacji

Do wykrywania ciągłości informacji CI posłużymy się lokalnym podobieństwem informacji dziennej w sąsiadujących dniach. Do tego celu wykorzystamy wcześniej zdefiniowane $\mathcal{E}(\mathcal{T}_i(d))$ w definicji trójek (10) oraz miarę podobieństwa kosinusowego:

$$
\begin{aligned}
cos(\theta) &= \frac{\mathcal{E}_1 \cdot \mathcal{E}_2}{||\mathcal{E}_1||\,||\mathcal{E}_2||} \\
cos(\theta) &= \frac{\mathcal{E}(\mathcal{T}_i(d_z)) \cdot \mathcal{E}(\mathcal{T}_j(d_{z+k}))}{||\mathcal{E}(\mathcal{T}_i(d_z))||\,||\mathcal{E}(\mathcal{T}_j(d_{z+k}))||} \\
k &= \begin{cases} +1 & \text{gdy jutro} \\ -1 & \text{gdy wczoraj} \end{cases}
\end{aligned}
\tag{11}
$$

w której $\mathcal{E}_i$ reprezentują embedingi porównywanych ze sobą $\mathbb{DI}_i(d_z)$ oraz $\mathbb{DI}_j(d_{z+/-1})$:

$$\begin{aligned} \mathcal{E}_i &= \mathcal{E}(\mathcal{T}_i(d_z)) \\ \mathcal{E}_j &= \mathcal{E}(\mathcal{T}_j(d_{z+k})) \end{aligned}$$
$$cos(\mathcal{E}_i, \mathcal{E}_j) = \frac{\sum_{l=1}^{1024} \mathcal{E}_{i_l}\, \mathcal{E}_{j_l}}{\sqrt{\sum_{l=1}^{1024} \mathcal{E}_{i_l}} \cdot \sqrt{\sum_{l=1}^{1024} \mathcal{E}_{j_l}}} \tag{12}$$

3.3.2 Definiowanie ciągłości

Dla każdej informacji w dniu z z reprezentacją $\mathcal{E}_i$ obliczane jest podobieństwo zgodnie z (12) do dnia *jutrzejszego* $z+1$ i *wczorajszego* $z-1$ z reprezentacją $\mathcal{E}_j$. Obliczając podobieństwo, wykorzystywane są pełne embeddingi o $l = 1024$ dla tekstowych reprezentacji $\mathcal{T}_i(d)$. W ten sposób powstaje macierz $\mathbf{MDI_f}$ podobieństw informacji w dniach sąsiadujących ze sobą:

$$\mathbf{MDI_f} = \begin{bmatrix} cos(\mathcal{E}_i, \mathcal{E}_j) : \forall \mathcal{E}_i \in \mathbb{DI}(d_z),\ \forall \mathcal{E}_j \in \mathbb{DI}(d_{z-1}) \\ cos(\mathcal{E}_i, \mathcal{E}_j) : \forall \mathcal{E}_i \in \mathbb{DI}(d_z),\ \forall \mathcal{E}_j \in \mathbb{DI}(d_{z+1}) \end{bmatrix} \tag{13}$$

po czym ograniczana jest ona tylko do podobieństw, które są powyżej zadanego progu podobieństwa β:

$$\begin{aligned} \mathbf{MDI_s} = \operatorname{argmax} \left(\left[x : \forall x \in \mathbf{MDI_f},\ x = \begin{cases} x & \iff x \geq \beta \\ 0 & \iff x \not> \beta \end{cases} \right] \right) \\ |\mathbf{MDI_s}| = i, i \in \langle 0, ..., 10 \rangle \\ \mathbf{MDI} = \mathbf{MDI}_{s_0} \times ... \times \mathbf{MDI}_{s_k} \end{aligned} \tag{14}$$

i to właśnie te newsy, które reprezentowane są przez niezerowe podobieństwa w pierwszym wierszu $\mathbf{MDI_s}$ wskazują w *Przeglądarce Informacji* na artykuły z podobnym, semantycznym ładunkiem informacyjnym z wczoraj. Zaś drugi wiersz w macierzy $\mathbf{MDI_s}$ wskazują na podobne informacje z *jutra*, które spełniły podstawowy warunek podobieństwa, czyli podobieństwo danej informacji w dniu do informacji z *jutra*, było większe od progu $\beta = 0.35$ jednak z ograniczeniem maksymalnie k najbardziej podobnych. Wartość β celowo została ustawiona na niską wartość, ponieważ z perspektywy przeglądarki, część artykułów pojawiających się w sąsiadujących dniach, przeplata informację między kolejnymi dniami z innymi informacjami. Warto wspomnieć, że liczba przykładów podobnych $i = 0$ możliwa jest wtedy, kiedy żaden $\mathcal{E}_j$ w dniu następnym/poprzednim nie spełnił warunku podobieństwa β do $\mathcal{E}_i$ w dniu *dzisiejszym*.

3.3.3 Ekstrakcja informacji ciągłej

Ekstrakcja informacji ciągłej bazuje na macierzy podobieństw **MDI** oraz obejmuje proces konwersji tej macierzy do grafu reprezentującego $\mathbb{DI}$ i relacje między

nimi. W tym procesie powstaje *skierowany* globalny graf informacji **GI**:

$$\begin{aligned} \mathbf{GI} &= \{\mathbf{V}, \mathbf{E}, \mathbf{W}\} \\ \mathbf{V} &= \{\mathbf{v}_0, ..., \mathbf{v}_i\} \\ \mathbf{E} &= \{\mathbf{e}_0, ..., \mathbf{e}_j\} \\ \mathbf{W}_{i\times i} &= \begin{bmatrix} w_{0,0} & \dots & w_{0,i} \\ \dots & \dots & \dots \\ w_{i,0} & \dots & w_{i,i} \end{bmatrix} \end{aligned} \tag{15}$$

w którym **V** reprezentuje zbiór węzłów tego grafu, **E** to zbiór krawędzi, a **W** to macierz wag do ekstrakcji informacji ciągłej. Równanie (15) przedstawia ogólną definicję globalnego **GI**. Proces powstawania węzłów **v** i krawędzi **e** w **GI** jest następujący:

$$\begin{aligned} &\forall\, \mathbb{DI}_i(d_j) \in \langle \mathbb{DI}_i(d_0), ..., \mathbb{DI}_j(d_k) \rangle \\ &\qquad \mathcal{M}_{\mathcal{V}}(\mathbb{DI}_i(d_j)) \xrightarrow[key]{unique} \mathbf{v}_m \\ &\qquad \mathcal{M}_{\mathcal{E}}(\mathbf{MDI_s}) \xrightarrow[key]{unique} \mathbf{e}_n,\ \mathbf{e}_n = (\mathbf{v}_i \to \mathbf{v}_j),\ i < j \end{aligned} \tag{16}$$

gdzie k to liczba analizowanych dni od dnia 0 w kolejności do dnia k-tego, $\mathcal{M}_{\mathcal{V}}$ to funkcja mapująca dowolne $\mathbb{DI}_i(d)$ na unikalny węzeł $\mathbf{v}_m$, a $\mathcal{M}_{\mathcal{E}}$ to funkcja mapująca podobieństwa informacji dziennej dla każdego $\mathbf{MDI_s}$ na krawędź e reprezentującą to podobieństwo w **GI**. Warto zaznaczyć, że i i j wskazują na pochodzenie węzła z informacji z różnych dni. Ważna jest również kolejność, zawsze w kierunku $i \to j$ lub $j \to i$. Nie zawsze, jeżeli istnieje $i \to j$ istnieje też $j \to i$ – wszystko zależy od wartości znajdujących się w $\mathbf{MDI_s}$. Podsumowując, Macierz $\mathbf{W}_{i\times i}$ zawiera wartości wag, które powstają podczas mapowania podobieństw z $\mathbf{MDI_s}$ na wagi każdej krawędzi $e_i \in \mathbf{E}$ za pomocą $\mathcal{M}_{\mathcal{M}}$:

$$\mathcal{M}_{\mathcal{M}}(\mathbf{MDI}_{s_i}) \to \text{tangens}(x) : \forall x_i \in \mathbf{MDI}_{s_i} \tag{17}$$

wykorzystanie w tym przypadku *tangensa* ma na celu zwiększenie odległości między niskimi, a wysokimi wartościami podobieństw z $\mathbf{MDI}_s$. Czyli konkretne $w_{i,j}$ to de facto $tan(cos(\mathcal{E}_i, \mathcal{E}_j)$, które przeszły warunek podobieństwa β na równaniu (14). Specyfika *tangensa* powoduje, ze niskie wartości podobieństwa zostaną przesunięte niewiele w górę, zaś im bliżej wartości $\beta = 1.0$, tym różnice stają się większe – te wartości znajdują się właśnie w macierzy wag $\mathbf{W}_{i\times i}$ i przypisane są do zbioru krawędzi **E**, gdzie:

$$\forall e_i \in \mathbf{E}\ \exists\, w_{k,l} \in \mathbf{W} \tag{18}$$

Posiadając zdefiniowane **GI** możemy rozpocząć opis *procesu ekstrakcji informacji ciągłej*, który w pseudokodzie przedstawiony został na **Algorytmie** (1). Po wykonaniu kroków z algorytmu (1), **GIF** zawiera wszystkie lokalne, ciągłe grafy informacyjne $\mathbf{GI}_i$. Proces ekstrakcji bazuje na losowym przechodzeniu przez graf, na początku ustalana jest losowa kolejność odwiedzanych węzłów,

Algorithm 1 Ekstrakcja $\mathbb{CI}$ (również $\mathbb{NCI}$)

Require: $\mathbf{GI} = \{\mathbf{V}, \mathbf{E}, \mathbf{W}\}$
Ensure: $|\mathbf{V}| \neq 0,\ |\mathbf{E}| \neq 0, |\mathbf{W}| \neq 0$

```
V' ← shuffle(V)                          ▷ V' - zbiór węzłów z losową kolejnością
R ← ∅                                    ▷ R - zbiór przeanalizowanych węzłów
GIF ← ∅                                  ▷ GIF - zbiór wszystkich grafów informacji ciągłej
while R ≠ V do
    v_f ← random(V')                     ▷ v_f - losowo wybrany węzeł z V'
    if v_f ∈ R then
        continue
    else
        R ← R ∪ {v_f}                    ▷ aktualizacja R o wybrany węzeł
    end if
    GI_i ← {v_f}                         ▷ GI_i to graf informacji ciągłej
    PRED ← G_pred(v_f)                   ▷ PRED - krawędzie wchodzące do v_f
    SUCC ← G_succ(v_f)                   ▷ SUCC - krawędzie wychodzące z v_f
    L ← PRED ∪ SUCC                      ▷ L - następniki i poprzedniki v_f
    for v_t ∈ L do
        e ← (v_f → v_t)                  ▷ e - krawędź łącząca v_f z v_t
        w ← W[v_f][v_t]                  ▷ w - waga krawędzi e w W
        if w < α then
            continue
        else if v_t ∈ R then
            continue
        else
            R ← R ∪ {v_t}                ▷ aktualizacja R o v_t
        end if
        e_n = None                       ▷ e_n – nowa krawędź do GI_i
        if v_t ∈ PRED then
            e_n = (v_t → v_f)
        else
            e_n = (v_f → v_t)
        end if
        if v_t ∉ GI_i then
            GI_i ← GI_i + v_t            ▷ dodawanie v_t do GI_i
        end if
        if e_n ∉ GI_i then
            GI_i ← GI_i + e_n            ▷ dodawanie e_n do GI_i
            GI_i[e_n] ← w                ▷ ustawianie wagi grawędzi e_n
        end if
    end for
    GIF ← GIF ∪ GI_i                     ▷ odłożenie GI_i do GI
end while
```

następnie losowo wybierany jest węzeł, od którego inicjowany jest pełen proces ekstrakcji. Wybór węzła na początku nie musiałby być losowy, ponieważ podczas przechodzenia sprawdzane są takie same warunki w każdym miejscu, jednak dla określenia stabilności samego algorytmu zdecydowaliśmy się na inicjalizację losowym węzłem. Od losowego węzła przechodzenie odbywa się jednocześnie w dwóch kierunkach: wszerz do przodu po liście $SUCC$ oraz wszerz do tyłu po liście $PRED$. Tworzone są nowe, lokalne grafy informacji ciągłej $\mathbf{GI}_i$, które aktualizowane są o węzły i krawędzie spełniające warunek $\mathbf{W}(e) \geq \alpha$. Do ekstrakcji grafów informacji ciągłej $\mathbb{CI}$ przyjęliśmy wartość $\alpha = 1.0$, co daje kompromis ekstrakcji informacji między bardzo szczegółową, a dość ogólną. Zwiększenie α spowoduje łączenie ze sobą informacji, które są do siebie bardziej podobne, zaś zmniejszanie będzie łączyło informacje mniej do siebie podobne. Proces budowy głównego grafu informacyjnego bazuje na analizie danych od 01.01.2025. Sam proces ekstrakcji nie jest również kierowany w kontekście konkretnych informacji – są to procesy nienadzorowane.

Podsumowując, można powiedzieć, że $\mathbb{CI}_i$ to informacja przedstawiona za pomocą $\mathbf{GI}_i$, która bez określania kierunku ekstrakcji, bazując na informacjach o podobieństwach lokalnych $n(d)$ między kolejnymi d, tworzy siatkę semantycznych powiązań, która następnie z globalnego $\mathbb{GI}$ ekstrahowana jest do mniejszych $\mathbb{GI}_i$ z zależnością występowania dni: $d \to d+1$ oraz $d-1 \to d$. Ważny jest kierunek przepływu, zawsze od dnia wcześniejszego, do dnia następnego, dlatego $\mathbf{GI}_i$ to skierowane grafy z zachowaną kolejnością występowania d i wagami z $\mathbf{W}$ przypisanymi do krawędzi e.

3.4 Informacja nieciągła $\mathbb{NCI}$

Aby otrzymać informacje nieciągłe, należy połączyć ze sobą informacje ciągłe. Dlatego podstawą definicji informacji nieciągłej $\mathbb{NCI}$ jest informacja ciągła $\mathbb{CI}$, a właściwie grafy informacyjne $\mathbf{GI}_i \in \mathbf{GIF}$ z algorytmu (1). Liczba informacji nieciągłych to ~ 600 grafów informacyjnych, a każdy z nich opisuje k newsów $n(d)$ z różnych dni d. Z poziomu $\mathbf{GI}_i$ możemy odczytać wszystkie n, które znajdują się w konkretnym $\mathbf{GI}_i$. Dlatego wracamy do definicji (10) i do tego celu, po skróceniu otrzymujemy zależność wyprowadzoną z wcześniejszych definicji:

$$\vec{N}_i(d) \,\forall\, \mathbb{DI}_i(d) \;\in\; \mathbf{GI}_i \tag{19}$$

którą wykorzystamy do określenia podobieństwa między $\mathbb{DI}_i(d_j)$ a $\mathbb{DI}_l(d_k)$. Grafy $\mathbf{GI}$ są dość mało zróżnicowane pod względem typów krawędzi (relacji między $\mathbb{DI}$). Dlatego pierwszym krokiem, jest ich przekształcenie do reprezentacji wektorowych, które ze sobą będzie można porównać aby określić $sim(\mathbf{GI}_i, \mathbf{GI}_j)$, do tego celu wykorzystamy zależność (19) i wyprowadzimy wzór na uśrednioną

wartość embeddingu $\mathcal{E}(\mathbf{GI})$ dla dowolnego $\mathbf{GI}$:

$$\begin{aligned} \mathcal{E}(\mathbf{GI}_z) &= \frac{1}{k}\sum_{i=1}^{k} \vec{N}_i(d)\, \forall \vec{N}_k \ \in \ \mathbf{GI}_z \\ \mathbb{E} &= [\mathcal{E}(\mathbf{GI}_i)\, \forall\, \mathcal{E}(\mathbf{GI}_z) \ \in \mathcal{E}(\mathbf{GI})] \\ &= \left[\mathcal{E}(\mathbf{GI}_1), ..., \mathcal{E}(\mathbf{GI}_{|\mathbf{GI}|}\right], \ |\mathbb{E}| = |\mathbf{GI}| \sim 600 \end{aligned} \tag{20}$$

W tej chwili macierz $\mathbb{E}$ z równania (20) zawiera uśrednione wartości z funkcji $\mathcal{E}$ mapującej newsy n na ich wektory $\vec{n}$ ze wszystkich dni d, które znalazły się we wszystkich $\mathbf{GI}_i \in \mathbf{GIF}$. Dzięki temu, możemy zastosować iteracyjnie mechanizm ekstrakcji informacji $\mathbb{CI}$ i budowy gragu $\mathbf{GI}$ opisanego na równaniu (15). Należy jednak wprowadzić kilka zmian w definicjach $\mathbf{V}$, $\mathbf{E}$ oraz $\mathbf{W}$. Budowany graf informacji nieciągłej oznaczymy jako $\mathbf{NGI}$, który budowany jest według następującego schematu:

$$\begin{aligned} \mathbf{NGI} &= \{\mathbf{V}, \mathbf{E}, \mathbf{W}\}, \ \mathbf{NGI} \subset \mathbf{GIF} \\ \mathbf{V} &= \{\mathbf{v}_0, ..., \mathbf{v}_i\}, \ \mathbf{v}_i \in \mathbf{GI}_j \\ \mathbf{E} &= \{\mathbf{e}_0, ..., \mathbf{e}_j\}, \ \mathbf{E} \ = \ \mathbf{V} \times \mathbf{V} \\ \mathbf{W}_{|\mathbf{V}|\times|\mathbf{V}|} &= \begin{bmatrix} w_{0,0} & \dots & w_{0,|\mathbf{V}|} \\ \dots & \dots & \dots \\ w_{i,0} & \dots & w_{|\mathbf{V}|,|\mathbf{V}|} \end{bmatrix} \end{aligned} \tag{21}$$

czyli węzeł w w grafie $\mathbf{NGI}$ jest grafem $\mathbf{GI} \in \mathbf{GIF}$, który reprezentowany jest za pomocą $\mathcal{E}(\mathbf{GI})$, zaś waga $w_{i,j}$ w macierzy $\mathbf{W}$ to podobieństwo między i-tym, a j-otym wektorem z $\mathbb{E}$. Do określenia wag w wykorzystany został mechanizm mapowania z równania (17) z przedefiniowaną funkcją mapującą węzły i krawędzie z definicji (16) tak, aby uwzględniały wcześniej wspomniane zmiany. Na tak zbudowany graf $\mathbf{NGI}$ (czyli bardzo gęsty graf z zależnościami $\mathbf{V} \times \mathbf{V}$) nakładany jest proces ekstrakcji grafu uogólnionego. Do tego celu wykorzystaliśmy algorytm (1) z ustawionym $\alpha = 1.2$ do ekstrakcji konkretnego $\mathbf{NGI}_i$, gdzie jako wynik otrzymujemy zależność z definicji (22).

$$\begin{aligned} \mathbf{NGI} \ &= \{\mathbf{NGI}_0, ..., \mathbf{NGI}_k\} \\ \mathbf{NGI}_i \ &\subset \mathbf{GIF} \end{aligned} \tag{22}$$

Podsumowując w jednym zdaniu czym jest $\mathbf{NCI}$: to zbiór nieciągłych grafów $\mathbf{NGI}$ składa się z konkretnych $\mathbf{NGI}_i$, które scalają w jednym grafie różne $\mathbf{GI}_j$, które składają się z różnych informacji $\mathbb{DI}$ składających się z różnych newsów n z różnych dni d. Innymi słowy, są to połączone grafy reprezentujące ciągłe informacje, do jednego grafu, w którym węzły są *grafami ciągłymi*, a krawędzie odzwierciedlają podobieństwo uśrednionych reprezentacji wektorowych grafów ciągłych, po ekstrakcji za pomocą algorytmu (1) wybranych przy założeniu $\alpha = 1.2$ do połączenia $\mathbf{GI}_i$ z $\mathbf{GI}_j$ do konkretnego $\mathbf{NGI}_k$.

4 Logout

Definicje przedstawione we wcześniejszych rozdziałach, to dokładne przedstawienie mechaniki i zasady działania *Przeglądarki* i *Kreatora Informacji* z naszego *playgroundu*. Prace przyszłościowe polegają głównie na koncentrowaniu się na *uśrednionych embeddingach* informacji nieciągłych, wykorzystaniu ich do przypisywania kategorii do pojawiających się newsów w czasie rzeczywistym. Planujemy oczywiście udostępnić embeddingi z opisem ich wykorzystania na naszym huggingface.