Co się pichci w polskim garnku?

Piotr Janyst
Mateusz Osikowicz
Jakub Licznerski
Wojciech Rauk

Wydział Informatyki i Zarządzania
kierunek informatyka, studia magisterskie
specjalizacja Danologia

Wprowadzenie

Polacy są zapalonymi kucharzami. Najpopularniejsze portale z przepisami zaliczają 8 milionów odwiedzin rocznie, a liczba użytkowników przekracza setki tysięcy. Na największych portalach z przepisami można znaleźć ponad 400000 przepisów, a wszystkie są opracowane przez dzielnych kucharzy w polskich domach.

W niniejszej pracy chcieliśmy zaprezentować opracowany **ważony graf powiązań składników przepisów**, które zbadaliśmy względem najpopularniejszych rodzajów kuchni.

Cel

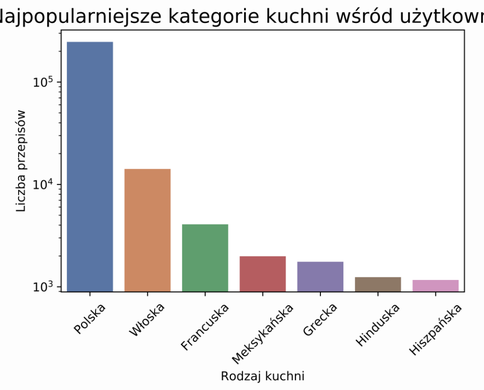
Wychowanie kulinarne i nawyki smakowe powodują, że potrawy z różnych krajów zawierają wiele Polskich elementów. Sprawia to, że definicja kuchni staje się bardzo złudna. Żeby wyciągnąć sedno kuchni wykonaliśmy skomplikowane przetwarzanie danych, a następnie wykorzystaliśmy metody analizy grafu do ekstrakcji esencji kuchni zagranicznych, eliminując wpływ Polskiej kultury kulinarnej.

Dane

Pobraliśmy dane z popularnych polskich serwisów z przepisami.

Nazwa kuchni	Liczba przepisów	Liczba składników
Kuchnia Polska	247107	2515
Kuchnia Grecka	1756	122
Kuchnia Włoska	14172	611
Kuchnia Francuska	4073	277
Kuchnia Hiszpańska	1164	83
Kuchnia Meksykańska	1985	161
Kuchnia Hinduska	1242	114

Najpopularniejsze kategorie kuchni wśród użytkowników



Najpopularniejsze składniki na polskich portalach z przepisami

Przygotowanie danych

Pobrane dane wymagały uważnego przetwarzania. Zaszumienie składników zostało wyeliminowane przy wykorzystaniu dwóch metod:

- Lematyzacji nazw składników
- Manualnemu czyszczeniu (np. usuwanie literówek, znaków specjalnych)

mąka,
mąki,
2 g mąki,
1 kg mąki,
mąkę,
mąka (ile zabierze),
0,5 kg mąki,
szkl. mąki,
mąka, 2 szklanki,
masło i mąka,
mąka (+do posypania blatu),
150g mąki,
250g mąki,
mąka na ciasto,
50g mąki,
mąka (kruszonka),
150 g mąki,
kruszonka; mąka,
3 g mąki,
mąka 1/2 szkl,
500 g mąki,
mąka na zasmażkę,
mąka 250 g,
mąka, 1 szkl,
...

→ **mąka**

Składnik polski	Składnik francuski	Miara podobieństwa
ser	ser pleśniowy	0.996470
parówki	krewetki	0.995904
brokuł	ryż	0.994236
sos sojowy	pomarańcza	0.993974
budyń waniliowy	cukier brązowy	0.991760

Składnik polski	Składnik grecki	Miara podobieństwa
olej do smażenia	jogurt grecki	0.952573
woda	feta	0.946215
olej	ogórek	0.941434
śmietana	cebula czerwona	0.939892
śmietana	papryka czerwona	0.936558

Składnik polski	Składnik włoski	Miara podobieństwa
czosnek ząbki	pomidory suszone	0.997926
parówki	rodzynki	0.997331
jogurt grecki	feta	0.996092
jogurt grecki	sos sojowy	0.995984
smalec	serek mascarpone	0.993515

W tabelach widać przykładowe rezultaty zastosowania miary ***similarity*** dla kuchni polskiej i wybranych. Ze względu na duży szum i niedoskonałość metody, nie otrzymano w pełni satysfakcjonujących wyników. Niektóre pary o wysokim podobieństwie mogłyby jednak być interpretowane jako, w uproszczeniu, substytuty. Zaznaczono je w czerwonej ramce.

Zbadaliśmy **polskie wyobrażenie** o **zagranicznych kuchniach**

W celu zdefiniowania **podstawowych różnic** pomiędzy badanymi kuchniami, znaleziono składniki których **częstotliwość występowania** jest większa niż **w kuchni polskiej** oraz dodano ich najbliższe sąsiedztwo:

Składniki definiujące kuchnie świata

Kuchnia	Najpopularniejsze składniki	Definiujące składniki
Indyjska	sól cebula curry czosnek pieprz	curry cebula pierś z kurczaka czosnek mleko kokosowe
Meksykańska	sól cebula papryka czerwona czosnek pieprz	papryka czerwona cebula czosnek fasola czerwona kukurydza
Francuska	sól jajka masło pieprz ciasto francuskie	ciasto francuskie masło oliwa tymianek jajka
Grecka	sól oliwa pieprz pomidor ser feta	oliwa ser feta pomidor czosnek ogórek
Hiszpańska	sól oliwa cebula jajka pieprz	oliwa pomidor papryka czerwona cebula czosnek
Włoska	sól oliwa pieprz cebula czosnek	oliwa czosnek pomidor oregano parmezan

Do wizualizacji podgrafu kuchni wybrano **5 składników** definiujących, a następnie dobrano ich **najsilniej powiązanych sąsiadów** (względem wagi relacji). Wielkość węzłów została ustalona na podstawie **stopnia węzła** (liczby powiązań). Szerokość połączenia definiuje **waga relacji**.

Tworzenie grafu

Z zebranych **377881 przepisów** w języku polskim, z których wykorzystaliśmy dostępne informacje do oznaczania kategorii kuchni składników w nich występujących. Z wszystkich przepisów wydzielono prawie 48925 unikalnych nazw składników. Blisko 100000 przepisów nie było oznaczonych żadną kategorią kuchni.

Liczenie wag θ w grafie

Jako wagę składnika przyjęto iloraz jego liczby wystąpień i ilości przepisów badanej kuchni.

$$\theta_{a,K} = \frac{N_{a,K}}{N_K}$$

Jako wartość połączenia pomiędzy dwoma składnikami zdecydowaliśmy się na miarę, która bierze pod uwagę liczbę współwystąpień badanych składników i wystąpień tych składników niezależnie od siebie.

$$\theta_{a\leftrightarrow b,K} = \frac{W\ (a,b)}{(N_a + N_b)\ * N_K}$$

gdzie:
$N_{a,K}$ – ilość wystąpień składnika ***a*** w przepisach kuchni **K**
N_K – ilość przepisów w jednej z siedmiu badanych kuchni
$W\ (a,b)$ – ilość współwystąpień składników ***a*** i ***b***

Po wyliczeniu wszystkich wartości wagi zostały znormalizowane do przedziału $\theta \in [0,1]$,

Potok przetwarzania danych

Pobieranie danych z popularnych polskich stron z przepisami — Cassandra, Flask, MongoDB
↓
Przetworzenie i normalizacja danych — Python, pandas, Morfeusz2
↓
Potok przetwarzania danych do postaci grafu — Apache Spark
↓
Udostępnienie powstałego grafu — AWS, Docker, Kubernetes, Neo4j

Metody analizy

- **Wagi węzłów i relacji** - wartości te symbolizują częstość występowania i współwystępowania składników i są podstawowym aspektem analizy.
- **Centrum (ang. *hub*)** - węzły o dużej ilości połączeń symbolizują składniki uniwersalne zarówno na przestrzeni wszystkich jak i w konkretnej kuchni, miara ta nie daje nam jednak wiele w analizie porównawczej na przestrzeni różnych kuchni ze względu na brak normalizacji i liniowy charakter rozkładu stopni grafu.
- **Betweenness** – pozwala znaleźć składniki spajające inne składniki w ramach kuchni, o centralnej pozycji w grafie. Podobnie jednak jak w przypadku centrum nie jest wartościowe w analizie porównawczej.
- **Similarity (Euclidean distance)** – w badaniach podobieństwa kuchni posłużono się miarą odległości między węzłami zdefiniowaną jako odległość euklidesowa. Liczona dla 50 najbliższych sąsiadów wzorem $w_{ab} = \theta_b * \theta_{a\leftrightarrow b}$ definiuje osadzenie węzła w przestrzeni wag. Wyliczonej dla każdej pary wierzchołków między grafami dwóch kuchni pozwala na znalezienie odpowiedników. Niestety metoda ta jest podatna na szum i jej wyniki musiały być przetworzone.

Wyniki i wnioski

⇒ Łatwość wyciągania ciekawostek: w kuchni indyjskiej nie występuje wołowina, a w Polskiej nie występują małże.

Użytkownicy portalów z przepisami posługują się niejednolitymi wariantami notacji składników. ⇐

⇒ Skompresowaliśmy liczbę składników o 1600 %.

Przy analizie similarity można zauważyć substytuty składników pomiędzy kuchniami. ⇐

⇒ Istnieje wiele składników takich jak sól, pieprz, jajka, które są często powielanymi składnikami pomiędzy różnymi kuchniami.

Separacja składników definiujących na podstawie różnicy wag z kuchnią polską działa bardzo dobrze. ⇐

⇒ Kuchnie po normalizacji mają zauważalnie różne składniki definiujące.

Cebula jest najbardziej dominującym warzywem w interpretacjach zagranicznych przepisów. ⇐

⇒ Przepisy w kuchniach są niezbalansowane względem kategorii. Wśród polskich potraw występuje wiele deserów.

Dalsze kroki i możliwe wykorzystania

- Pomimo wyraźnej dominacji liczności przepisów w kuchni Polskiej udało się wyodrębnić składniki definiujące najpopularniejsze zagraniczne kuchnie w Polsce.
- Możliwe zwiększenie wartości informacyjnej grafu można byłoby osiągnąć poprzez wprowadzenie hierarchii składników. Tak aby móc wykorzystać semantyczną różnicę, ale i związek między na np. serem żółtym a pleśniowym.
- Utworzony graf, rozszerzony o odpowiednie metadane, może być pomocny przy tworzeniu diet z braniem pod uwagę upodobań kulinarnych użytkownika jak również do budowania reprezentacji liczbowej produktów spożywczych.

www.food-graph.tech