VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ FAKULTA INFORMAČNÍCH TECHNOLOGIÍ

Metavyhledávání recenzí na českém webu

Autor - Šimon Matyáš - xmatya11@stud.fit.vutbr.cz
Vedoucí - doc. prof. RNDr. Pavel Smrž, Ph.D.

Úvod

Je mnoho novinkových webů a blogů zabývajících se recenzováním různých produktů. Tyto recenze se ovšem často k uživateli vůbec nedostanou, pokud nehledá právě recenze konkrétního produktu.

Uživatelé k ohodnocení produktu využívají krátké neprofesionální recenze od ostatních uživatelů, kteří si daný produkt zakoupili před krátkou dobou.

Českých portálů píšících rozsáhlejší recenze je mnoho, ovšem většinou jsou zaměřeny pouze na jeden produkt typ produktu (svetandroida.cz, auto-mania.cz).

Práce se tedy hlavně zaměřuje na vyhledávání článků obsahující rozsáhlejší recenze, které uživateli poskytnou objektivnější ohodnocení produktu.

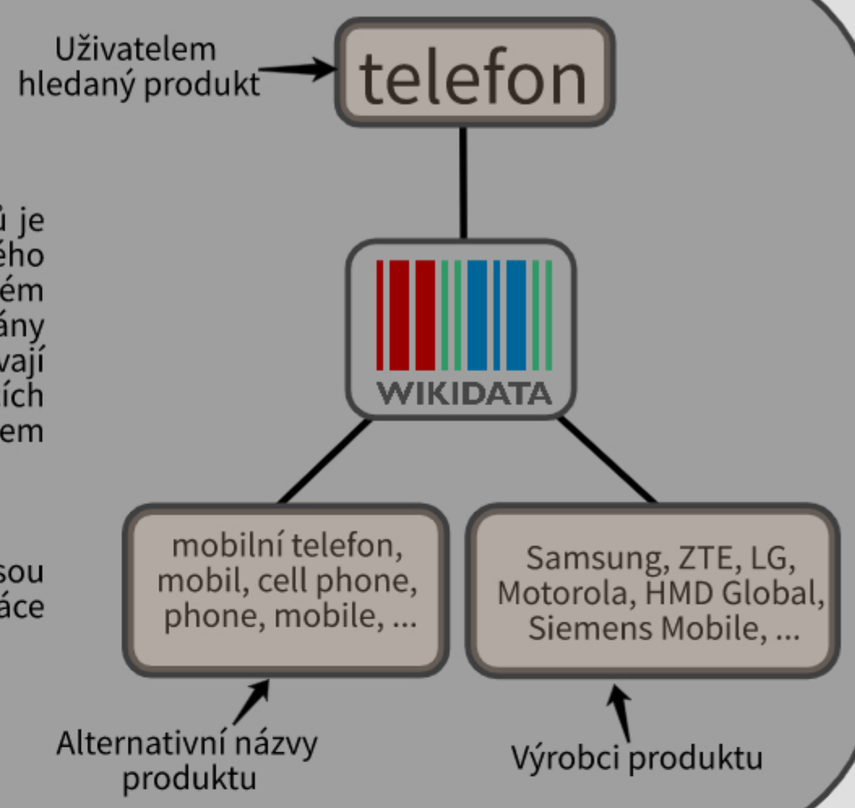
Práci mohou využít i pracovníci portálů shromažďující recenze k vyhledání recenzí.

Cíl práce je tedy vytvořit systém který vyhledá články obsahujících recenze na zadanou skupinu produktů, jako jsou například "telefon", "monitor", "auto" a je schopen vyhledat i věci, jako jsou třeba "film", "kniha" a "počítačová hra". Dále vyhledané recenze ohodnotí jestli článek hodnotí produkt pozitivně či negativně.

Vyhledání alternativních názvů produktů a výrobců

Pro vyhledávání celé skupiny produktů je třeba najít alternativní názvy zadaného produktu které se mohou v hledaném článku vyskytnout, navíc jsou vyhledány firmy které se daným produktem zabývají protože jsou většinou v článcích zabývajících se hledaným produktem zmíněny.

K vyhledávání těchto informací jsou použity Wikidata ve kterých práce vyhledává pomocí jazyku SPARQL.

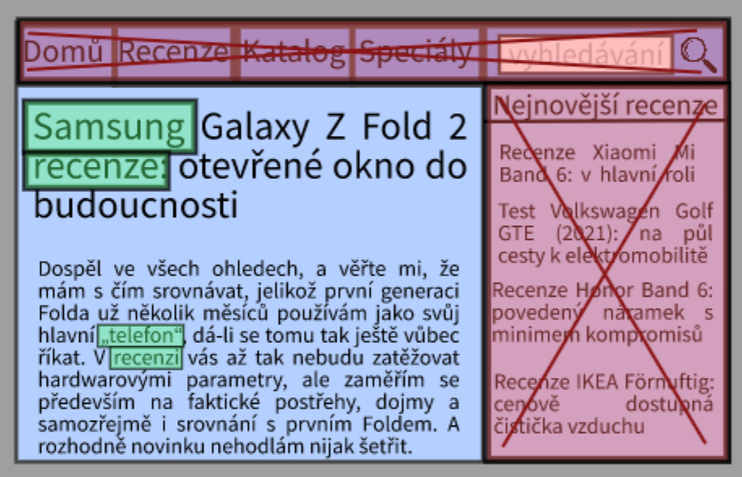


Určení recenze

Pro správné rozpoznání recenze je třeba vybrat pouze nadpisy a obsah článku kvůli tomu že se v textu často vyskytují odkazy na podobné články nebo je někde na stránce sekce "Mohlo by vás zajímat" která obsahuje mnoho dalších odkazů. Práce se tedy snaží vybírat pouze relevantní text.

Recenze jsou rozpoznány podle slov v relevantním textu článku, jednotlivá slova textu jsou porovnávány se seznamy slov označující recenze, názvy hledaného produktu a výrobci daného produktu. Pokud je základní tvar slova v některém seznamu z hledaných slov je článku přičteno skóre které znázorňuje podobnost s textem obsahující recenzi hledaného produktu.

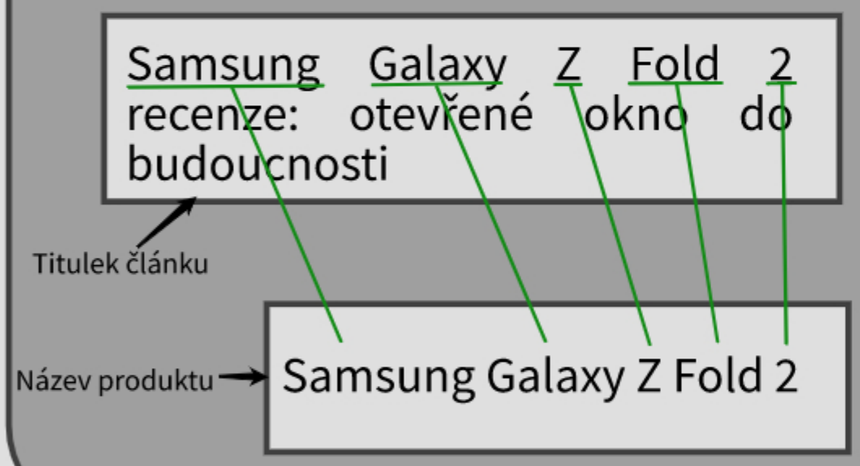
Hodnota která je článku přičtena při nalezení slova v některém ze seznamů záleží na několika aspektech, jako ve kterém ze seznamů bylo slovo nalezeno a taky kde v textu se toto slovo nachází.



Určení názvu produktu

Název produktu kterým se článek zabývá je identifikován pomocí extrakce slov z titulku. Jsou vybrány slova začínající velkým písmenem či číslovkou.

Pro případy kde se název produktu nedá rozpoznat na základě velkých písmen jsou vytvořeny speciální případy pomocí kterých je název produktu zjištěn extrakcí z HTML kódu.

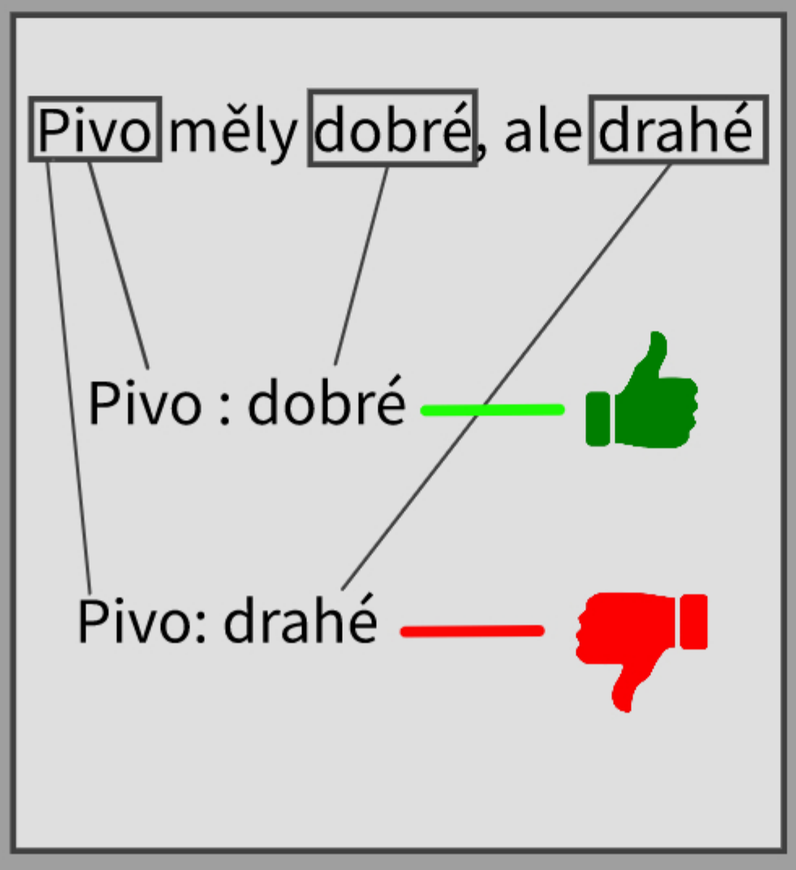


Určení sentimentu

U článků je ohodnocen sentiment autora recenze, hlavně jestli recenzent hodnotí produkt pozitivně či negativně. Primárně je sentiment určován pomocí aspektové analýzy, v textu zpracovávaného článku jsou vyhledány aspekty produktu a jejich hodnocení, poté je určeno jak jsou jednotlivé aspekty ohodnoceny.

Aspekty jsou v textu vyhledány na základě větné stavby která je vytvořena pomocí nástroje udpipe

Pokud se v textu nezdaří vyhledat žádné vhodné slova na základě kterých by šel aspekt ohodnotit je provedena analýza sentimentu pomocí modelu BERT společně s knihovnou od společnosti.



Závěr

Práce má za úkol vyhledat články obsahující recenze zadaného produktu na českých webech, rozpoznat o jakém produktu se v textu píše a určit sentiment obsahu článku. Tento úkol práce splňuje.

I přesto že systém správně vyhledá články obsahující recenze, ve výsledcích se vyskytují i špatně zvolené články, je tedy třeba v některých případech ručně vybírat které články jsou recenze a které ne. Výsledky jsou seřazeny podle relevantnosti.

Aktuálně je práce nahraná pouze na serverech společnosti KNoT, kde má přístup ke zdrojovým datům ve kterých jsou vyhledávány články. To znamená, že systém mohou použít pouze uživatelé kteří mají k tomuto serveru přístup, v budoucnu by mohlo být vytvořeno webové rozhraní aby práci mohli využít i uživatelé, kteří k tomuto serveru přístup nemají.

Práci mohou využít pracovníci portálů zaměřující se na recenze jako například heureka.cz na vyhledávání článků obsahující recenze a pak tyto články doplnit k produktům na portálu, pro který pracují.