Hľadanie anomálií v textoch

použitím hlbokých neurónových sietí

Vedúci: doc. RNDr. Gabriela Andrejková, CSc.

Autor: Bc. Zoltán Szoplák

Ciele práce:

1. Spracovať prehľad existujúcich metód používajúcich neurónové siete pre hľadanie anomálií v textoch.

2. Spracovať prehľad metód pre reprezentáciu textových dát s cieľom ich ďalšieho spracovania.

3. Pripraviť vhodné textové dáta na trénovanie a vyhodnotenie neurónovej siete.

4. Navrhnúť a implementovať hlbokú neurónovú sieť na vyhľadávanie outlierov v texte.

5. Vyhodnotiť efektivitu neurónovej siete a porovnať s výsledkami získanými inými metódami.

Rôzne texty akými sú knihy, časopisy, články alebo aj jednoduché správy či letáky disponujú vlastnou štruktúrou a charakteristikami. Vhodnou analýzou nejakej množiny smerodajných atribútov je možné z textových dát získať tzv. unikátny autorský štýl. Ak má text jediného autora, malo by platiť, že štýlové atribúty vzťahujúce sa na celý text by sa mali viac menej zhodovať so štýlovými atribútmi jednotlivých častí daného textu. Ak toto neplatí, vieme prehlásiť že v našom texte sa nachádzajú nejaké nezrovnalosti, anomálie(angl. outliers). Anomáliami v štatistike sa nazývajú také pozorovania, ktoré sa vôbec nezhodujú so štatistickým celkom, tvoreným ostatnými pozorovaniami. Nájdenie takejto anomálie v texte naznačuje, že do textu bolo zasiahnuté nejakou ďalšou osobou alebo pochádza z iného zdroja. Keďže okrem analyzovaného textu mnohokrát nemáme k dispozícii žiadne iné, aby sme ich porovnali alebo našli zdroj nezhodnej časti, musíme si vystačiť s takouto analýzou, ktorá používa jediný text, vnútri ktorého sa hľadajú nezhodujúce sa časti. V rámci tejto práce sa budeme zaoberať dlhšími textami, ktoré sú z analytického dôvodu iného charakteru ako tie kratšie (pozvánky, letáky, reklamy, články v časopisoch atď.), ktoré by sa analyzovali iným spôsobom.

Takáto analýza je vhodná na odhalenie plagiátorstva v prípadoch, kedy nemáme možnosť dáta alebo čas to porovnať s inými vzorkami. Má to zmysel taktiež pri textoch napísaných pred modernou dobou, či už pri verifikácii autorstva, odhaleniu nejakých nezrovnalostí alebo nájdenie modifikácií ktoré vyvrátia jeho vierohodnosť. Táto analýza môže byť tiež použitá na odstránenie anomálií ako úprava dátovej sady pred použitím nejakej inej dátovej analýzy.

Podobné problematiky sú riešené v oblasti štatistiky či strojového učenia rôznymi metódami. V oblasti neurónových sietí je riešení taktiež viacero. Populárnym riešením sú konvolučné neurónové siete, no problém bol taktiež riešený cez Kohonenove samoorganizačné mapy celkom úspešne.

Táto práca bude používať iný postup. Základom budú hlboké neurónové siete, ktoré majú vlastnosť že sa vyhýbajú problému tzv. preučenia ("overfittingu") siete pri veľkom množstve dát, keďže disponujú veľkým množstvom neurónov a váh, čím sa významne dá zväčšiť dátová sada, ktorá bude použitá pre naučenie siete. Naďalej, štandardné strojové učenie robí extrakciu čŕt manuálne, teda v príklade textov sa zvolí nejaká reprezentácia ako n-gramy, či histogramy, zatiaľ čo vhodne navrhnuté hlboké neurónové urobia takéto extrakcie automatickým spôsobom. Týmto sa môžeme vyhnúť strate dát, ak sa rozhodneme pre nejakú všeobecnú abstrakciu.

Ako základný návrh siete by sme použili model neurónovej siete zvaný HTM (Hierarchical Temporal Memory), ktorá napodobňuje štruktúru jedného bloku neokortexu. Samotná štruktúra HTM má tvar pyramídy tvorenej navzájom prepojenými, ktorá dostane na vstup dáta surovo bez akéhokoľvek predspracovania(konkrétne v binárnej postupnosti) a v každej vrstve extrahuje nejaké črty na základe predošlých, ktorých zložitosť sa zdola nahor zvyšuje. Ak si zoberieme zrak ako príklad, ako vstup zoberieme obraz vysvietení na retinu, a v každej vrstve neokortikálnej štruktúry získavame čoraz komplexnejšie črty, napríklad v prvej vrstve nájdeme hrany a rohy, v ďalšej vrstve z týchto čŕt získame kontúry a základné útvary, a v najvyššej vrstve už nájdeme komplexné objekty (človek, strom) ako aj ich relatívnu pozíciu. Ďalšou charakteristikou metódy HTM že simulujú pamäť a to v takom zmysle, že vyššie vrstvy si pamätajú extrahované vzory omnoho dlhšie ako tie nižšie, čo sa zhoduje s realitou (ak pozorovaná osoba otočí hlavou, vieme, že sa jedná o rovnakú osobu, ktorú sme spozorovali predtým, bez toho, aby sme ju museli znovu poznávať, alebo fakt, že je jednoduchšie naučiť sa nové slovo z jazyka, keď ho ovládame ako keď nie. Ako z týchto vlastností vidíme, model je možné použiť pri mnohých úlohách súvisiacich s hľadaním komplexných vzorov, medzi ktoré patrí aj tá naša. Keďže model je výpočtovo náročný plánuje sa použiť cluster počítačov ako aj optimalizácia či už na úrovni siete (redukcia/úprava na základe konkrétnych znalostí spracovania textu) alebo na úrovni zefektívnenia samotných výpočtov. Model budeme trénovať a vyhodnocovať na textoch s umelo vygenerovanými anomáliami, kde do originálneho textu pridáme obsahovo podobnú časť z iného zdroja, čo budeme považovať za anomáliu.

Literatúra:

1. Ian Goodfellow and Yoshua Bengio and Aaron Courville: Deep Learning, MIT Press, 2016
2. Ramakrishnan Kannan, Hyenkyun Woo, Charu C. Aggarwal, Haesun Park: Outlier Detection for Text Data : An Extended Version, 2017
3. Honglei Zhuang, Chi Wang, Fangbo Tao, Lance Kaplan, Jiawei Han : Identifying Semantically Deviating Outlier Documents, Proceeding of 2017 Conference on Empirical Methods in Natural Language Processing (September 2017)
4.Charu C. Aggarwal - Outlier Analysis, 2013
5. Subutai Ahmad, Alexander Lavin, Scott Purdy, Zuha Agha: Unsupervised real-time anomaly detection for streaming data. Neurocomputing, 262 (2017) 134–147