Spam Filtr

**Autoři: Monika Svobodová, Vít Nademlejnský**

**Úvod**

* naším úkolem bylo vypracovat spam filtr
* filtr měl rozeznávat, jestli je email spamem nebo hamem

**Princip našeho spam filtru**

* náš spam filtr dostal dva seznamy slov
* v jednom byla slova, která mají větší pravděpodobnost se vyskytnout ve spamu
* např. slova upozorňující na nízké ceny, slevy, něco zdarma, požadavky hesel
* ve druhém byla slova, která se podle nás vyskytují v hamech
* např. slangové výrazy nebo naopak velmi formální výrazy
* pro každý email náš filtr zjistil počet spamových a hamových slov
* z toho pak určil pravděpodobnost, jak moc je email spam nebo ham
* podle toho, která pravděpodobnost byla větší, tak určil email jako spam nebo ham

**Trénování filtru**

* filtr se učí tak, že vezme všechna slova z emailů a zároveň počítá jejich výskyt v emailech označených jako spam nebo jako ham pomocí souboru !truth.txt
* slova, která se vyskytovala častěji ve spam a zároveň vůbec v ham emailech, jsme zařadili do seznamu slov spamových
* naopak slova, která měla vysoký výskyt v hamových emailech a žádný ve spamových emailech, jsme zařadili do slov hamových
* tyto seznamy pak nahradily naše původní seznamy pro určení spam nebo ham emailů

**Výsledky**

* během našeho testování jsme různě zkoušeli měnit hodnoty pro průchod slov do seznamů slov spamových nebo hamových
* na trénovacích sadách jsme dosahovali skvělých výsledků – kolem 90%
* avšak pro sady v Brute jsme při skvělých výsledcích na trénovacích sadách, dosáhli katastrofických výsledků
* proto jsme byli nuceni nastavení změnit, abychom dosahovali vyšší úspěšnosti na hlavních testovacích sadách

**Rozdělení práce**Obsah obrázku stůl

Popis byl vytvořen automaticky

**Organizace práce**

* Domluva přes MS Teams
* Zasílání souborů přes email

**Závěr**

Pro nás oba to bylo nové, pracovat ve dvojici na jednom větším projektu. Nikdy předtím jsme nic takového v oboru IT nedělali. Ačkoliv je velká škoda, že jsme se za této nepříznivé situace nemohli sejít, domluva probíhala v pořádku. Práce na spam filtru nás velice bavila a věříme, že nás i obohatila o cenné zkušenosti. Nevíme, jak bychom projekt stihli odevzdat do předem stanoveného deadlinu, takže jsme rádi, že se deadline odložil.

**Použitá literatura/stránky**

<https://www.activecampaign.com/blog/spam-words>

<https://www.wallstreetenglish.com/blog/how-to-write-formal-emails-in-english/>