Zadání projektu

V rámci projektu budete zpracovávat přiložené textové soubory.

Úkoly:

- Načíst textový soubor data.txt,
- načíst textový soubor stop_words.txt,
- zpracovávat datový soubor data.txt po slovech,
- odfiltrovat slova s počtem znaků větším než 8 a menší než 4,
- odfiltrovat slova podle stop slov ze stop_words.txt,
- spočítat statistiky – nejfrekventovanější slovo a počet jeho výskytů, nejméně frekventované slovo a počet jeho výskytů, celkový počet slov po filtracích,
- výsledný algoritmus by měl být co nejefektivnější,
- výsledky časů zpracování jednotlivými algoritmy (CPU, GPU, ….) vykreslete do koláčového grafu (použijte knihovnu Matplotlib),
- všechny požadované výstupy vypište do konzole a výsledky srovnejte.

Budou vytvořeny 4 verze algoritmu:

1. CPU – jedno vláknový algoritmus
2. CPU – více-vláknový algoritmus
    - Zpracovávejte paralelně datový soubor.
    - Pro výpočet využijte všechna dostupná CPU jádra.
3. GPU verze
    - GPU neumí jednoduše pracovat se datovým typem string, proto zde bude odlišný způsob výpočtu. Slova z datového souboru i ze stop slov musí přemapovat na číselné hodnoty – vytvořte si pomocný slovník (pro data.txt), kde budete mít uloženo vždy název slova a k němu zvolte jeho číslo id, aby jste dokázali s pomocí id (integer) zpětně mapovat název slova. Dále takto přemapujte textové soubory na vstupní vektory typu int, které již můžete zpracovat na GPU.
    - Odfiltrování slov podle počtu znaků se nebude v provádět na GPU (moc složité), ale tuto filtraci proveďte před nahráním dat na GPU.
    - Odfiltrování podle stop slov bude provedeno na GPU.
4. Apache spark verze
    - Využijte všechna dostupná CPU jádra.