M2 UE Appli spécialisée pour le TALNE
(Séance 1 Analyse d'opinion)

Cette première séance de TP a pour objectif, la prise en main des données et d'un langage (de préférence interprété awk ou perl). Une mini documentation du langage awk se trouve dans le fichier awk.pdf. Un extrait du corpus de micro critiques (MC) de films issues du site vodkaster.com se trouve dans le fichier AppVk13.txt. Ce document contient 6999 MC qui sont structurées en plusieurs champs séparés par des tabulations de la façon suivante :

Id Pseudo Date Heure Post Titre Film Note Micro Critique …

```
X     horizonj    2013-08-21T14…    LittleMissSunshine      4     Une perle du cinéma indépendant        2005  2323  US    …
X     horizonj    2013-08-21T14…    Fast & Furious 6        1.5   Il va falloir penser à s'arrêter...  2013  222   US
```

L'identifiant est unique. L'échelle des notes comporte 10 barreaux de 0,5 à 5

Les questions auxquelles il faut répondre à l'issue de cette première séance sont les suivantes :

1. Représenter sous forme d'histogramme la distribution des notes. En déduire la note moyenne. La note 0 correspond à pas de note (les MC ayant une note nulle ou pas de note doivent être exclues). Combien y a-t-il d'avis positifs et d'avis négatifs, si on considère que les notes supérieures ou égales à 4 sont positives, et celles inférieures ou égales à 2.5 sont négatives (les neutres valant 3 ou 3.5 sont ignorées).

2. Quels sont les 10 films qui ont reçu le plus de critiques ? (calculer la note moyenne associée à chacun d'eux)

3. Quels sont les 7 films les plus appréciés et les 7 films les moins appréciés selon un critère de votre choix ? (par exemple : en fonction des notes attribuées)

4. Indiquez comment sont réparties les notes associées aux MC contenant l'expression ***excellent film***. Même question avec l'expression ***mauvais film***.

5. Considérez chacune de ces expressions comme une requête et recherchez les critiques qui les contiennent. Pour chaque expression on disposera de plusieurs mesures : la note moyenne qui lui est associée, la dispersion, et enfin on détaillera la distribution de ces notes. Commentez ces statistiques.

DIVERS :

`=========`

Une "stop liste" mise à disposition par J. Véronis à l'adresse http://sites.univ-provence.fr/veronis/data/antidico.txt
Il sera peut-être utile de la modifier pour tenir compte des particularités de la tâche.

`==================`

Les étudiants s'engagent à ne pas conserver à la fin des TP les données (MC) qui sont mises à leur disposition dans le cadre de ce cours

Avignon, le 20 novembre 2013 Enseignant : Marc El-Bèze marc.elbeze@univ-avignon.fr