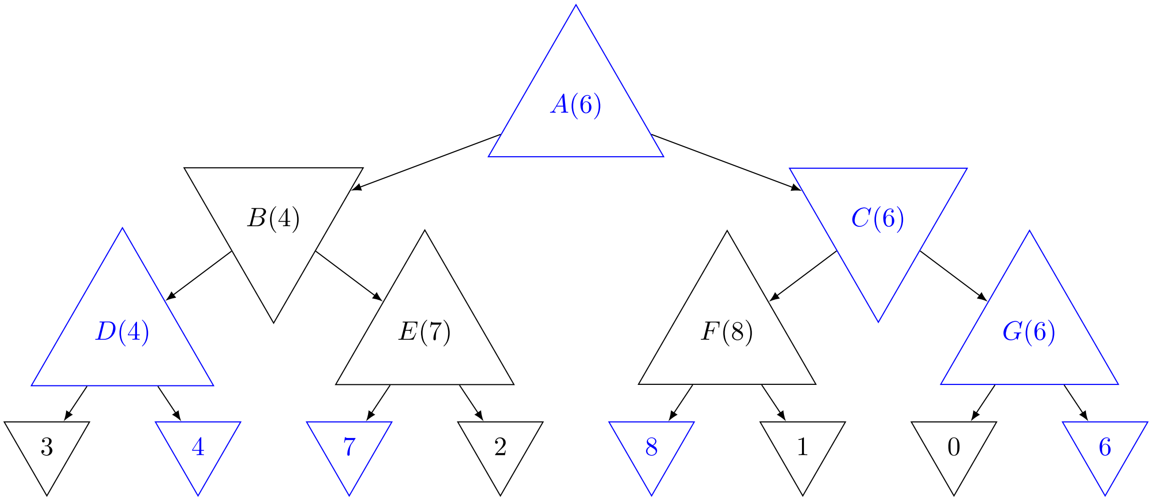
CS 540: HW5
Jurijs Nazarovs

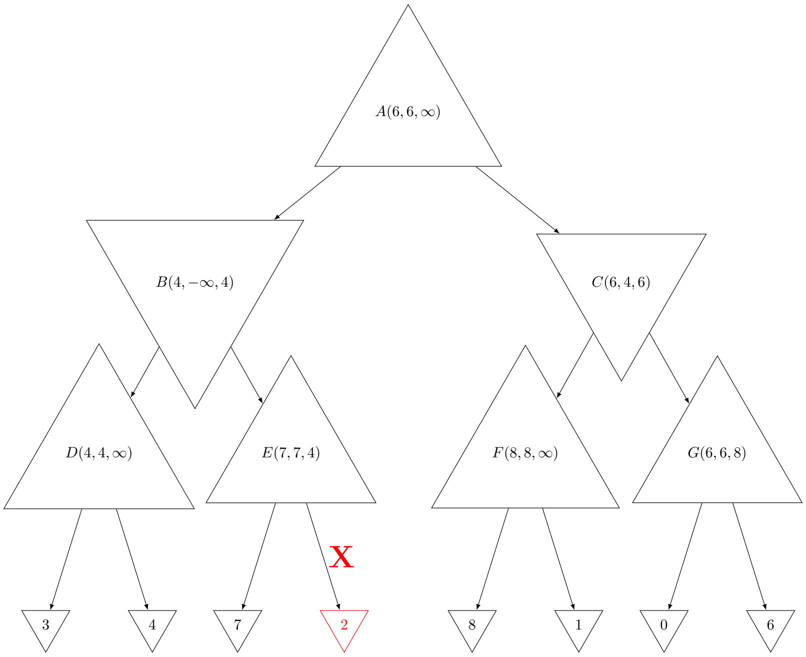
Question 1: Game Tree Search [60 points]

1. Minimax algorithm on this game tree



2. Alpha-beta pruning to compute the Minimax value at each node for the original game tree

   Order of numbers in node is (theoretical value, α, β).



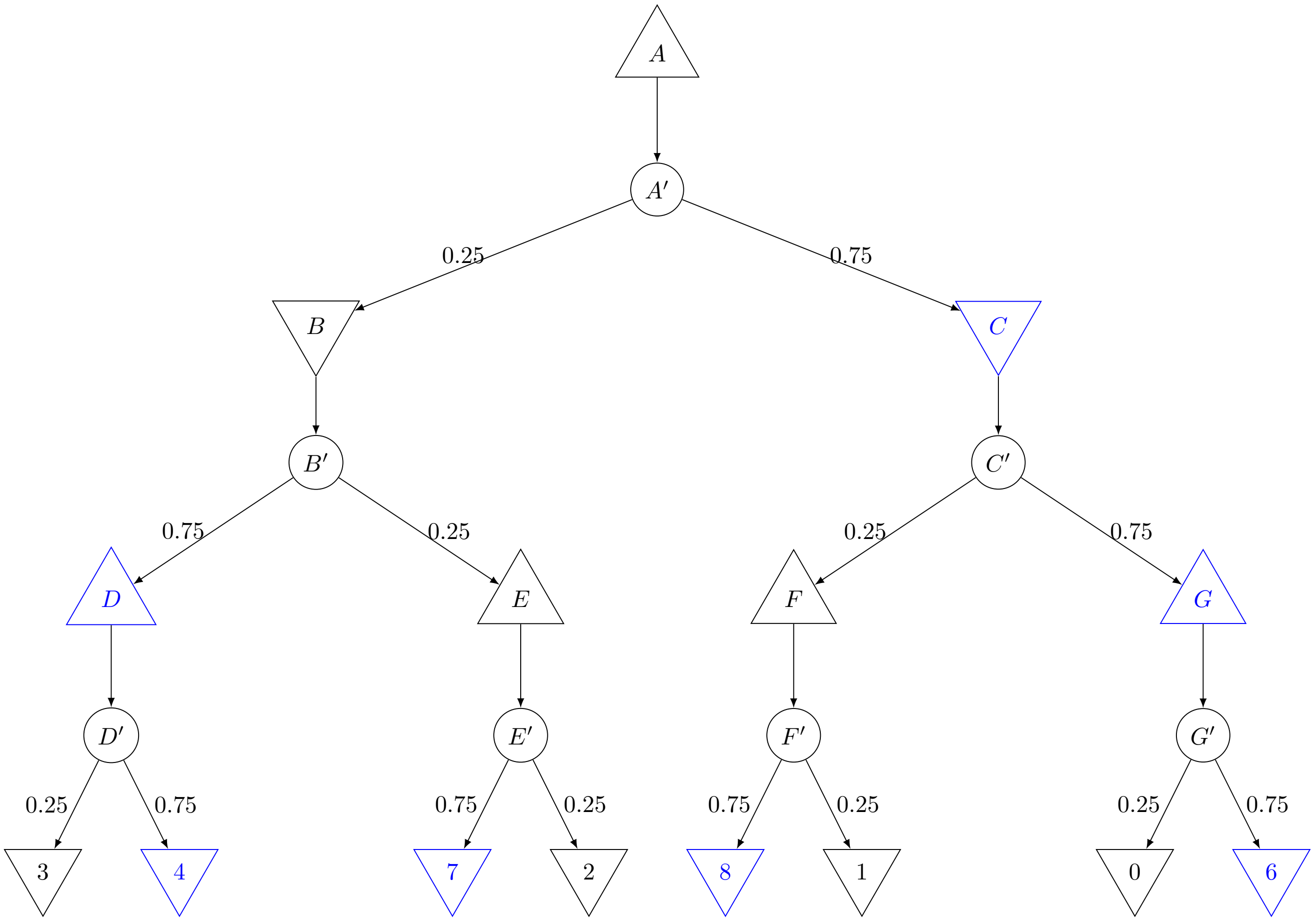
3. Pruned branches Pruned branches/nodes are highlighted with red color on a graph above. Namely, it is a node with value 2, child of node E.

4. Why we would want to use alpha-beta pruning?

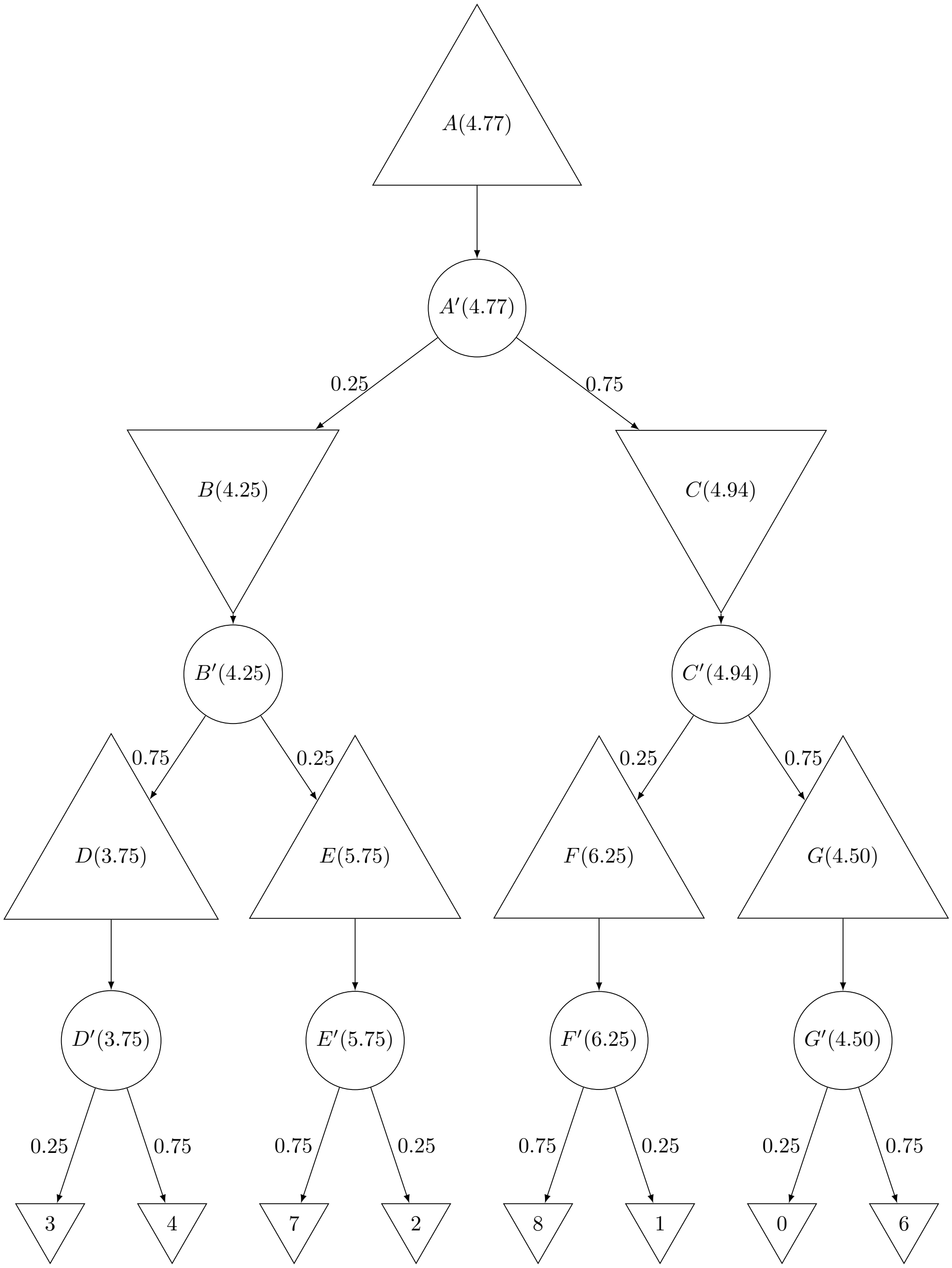
   We would want to use alpha-beta pruning to avoid going threw branches, which cannot make situation better. Such approach saves time in a tree analysis.

5. Probability of a player choosing an optimal action at each node? At every node the probability to choose an optimal action is $P = 0.5 + 0.5^2 = 0.75$.

6. Chance node



7. Non-deterministic game. Expected value is provided in the brackets for every node.



Question 2: Natural Language Processing [40 points, 5 points each]

1. I believe that the most convenient way to break the corpus into “computer words” is to use bash or analogue languages, since there is a lot of tools to work with text-processing. We can simply concatenate all files inside a directory, translate spaces into “new line” character, and proceed with sorting and unique commands using pipeline. Thus, as a word we consider a sequence of characters between spaces.

   One - line command, which takes just 0.38s, is:

   ```
   cat "$dir"/* | tr [:space:] "\n" | grep ’[^[:blank:]]’ | sort | uniq -c | sort -bgr > sorted.list
   ```

   Such program allows to add additional editing features, by applying more commands in pipeline. For example, we can make all words not case sensitive, or delete commas and other signs by adding command tr, e.g. tr “[:lower:]” “[:upper:]”.

2. Total words: 205057
   Unique words: 18383

3. Top 20 word types and their counts.

   ```
   10636 the
   6686 to
   6128 of
   5157 and
   3846 in
   3702 a
   3354 that
   3312 is
   2663 be
   2567 AI
   2022 will
   1615 for
   1613 are
   1517 it
   1505 not
   1483 on
   1483 as
   1221 with
   1140 The
   1066 have
   ```

4. Pick 20 bottom word types and their counts.

   ```
   1 ”NLP
   1 ”NHE-Fact-Sheet.”
   1 ”Mcanique
   ```

```
1 ”Meditations
1 ”Mars”.
1 ”Large-scale
1 ”Given
1 ”Ethical
1 ”Current
1 ”Competition
1 ”Changes
1 ”Bad
1 ”Back
1 ”Autonomous
1 ”As
1 ”ARTIFICIAL
1 ”A
1 ”2016
1 ”..on
1 ”...
```

5. Plot r on the x-axis and c on the y-axis, namely each (r_i, c_i) is a point in that 2D space (you can choose to connect the points or not).

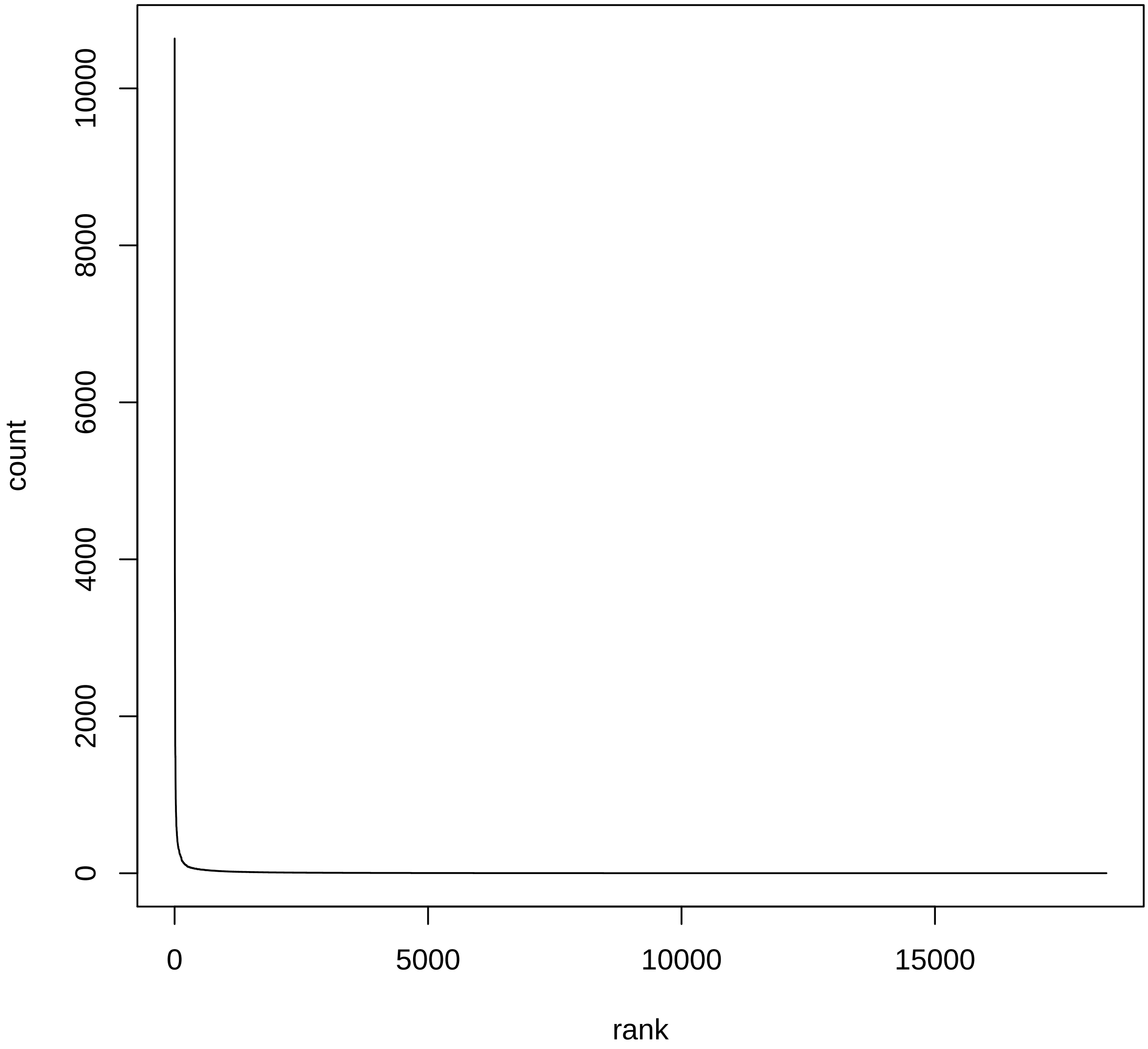


Figure 1: No scaling

6. Plot $\log(r)$ on the x-axis and $\log(c)$ on the y-axis.

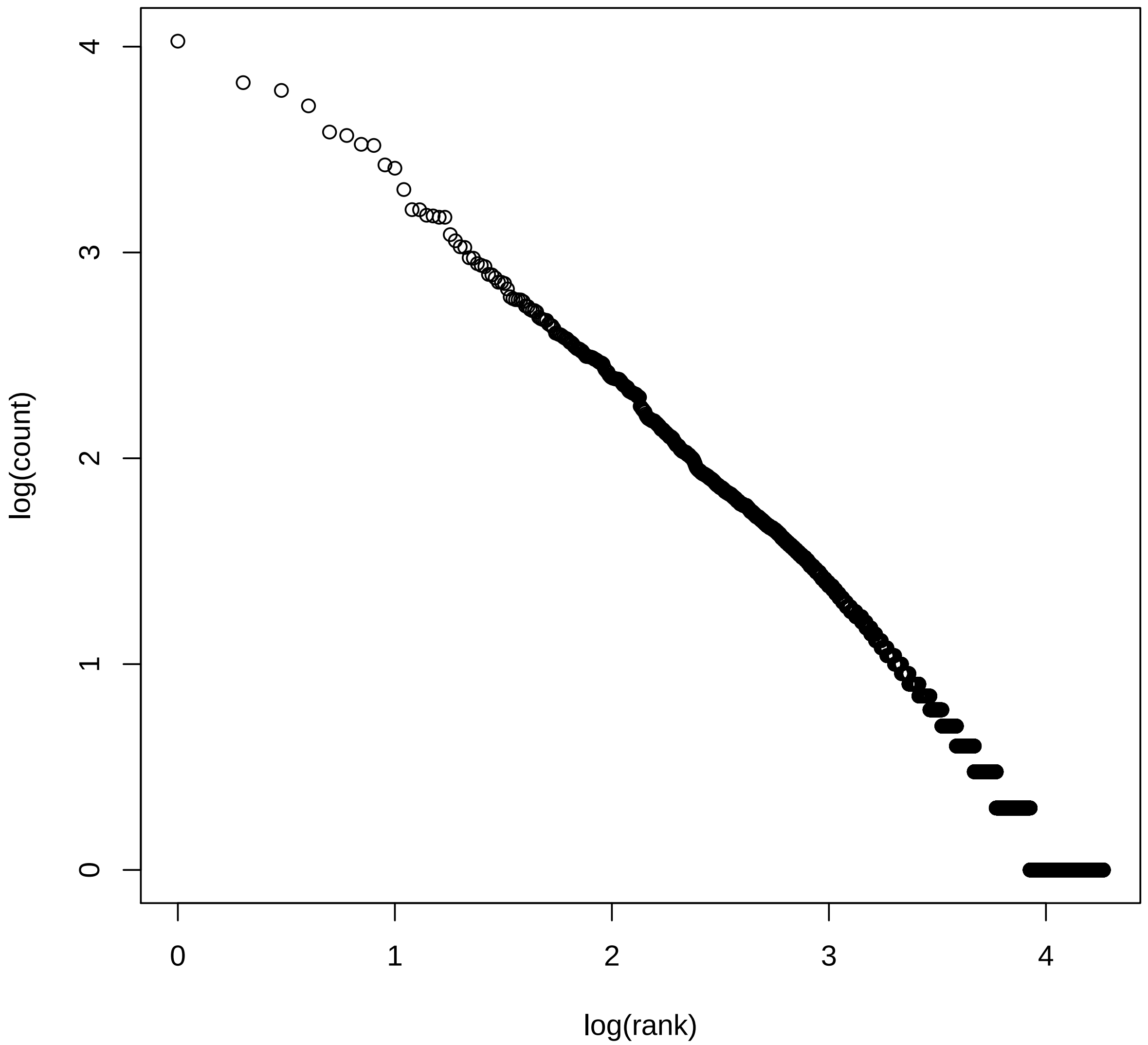


Figure 2: Log scaling for both variables

7. Briefly explain what the shape of the two curves mean.

   Both curves are non-increasing curves. Which shows that higher rank corresponds to lower count. From the plot of raw data, we can see that counts and rank are described by following relation: count $= \frac{1}{\text{rank}}$. Such observation is consistent with Zipf's Law, described in lectures as $f \propto \frac{1}{g}$.

Since from first picture we can assume that count $= \frac{1}{\text{rank}}$, then log(count) = −log(rank). That is why we see the negative slope of the graph on second image. And the graph looks more as a linear function rather the hyperbola.

8. Discuss *two* potential major issues with your computer words, if one wants to use them for natural language processing.

   1. Difference in character encoding standards of separate documents. I noticed that sometimes ' is presented as is, but sometime it is shown as $< 92 >$. Such issue leads to wrong results of counting values.

   2. Since I considered just a sequence of characters between spaces as a word, and not any of regular expressions, my program does not count for misspellings. Which might cause outliers in counts. Also, for the same reason, words combined with some symbols like " or commas are considered as different from just words.