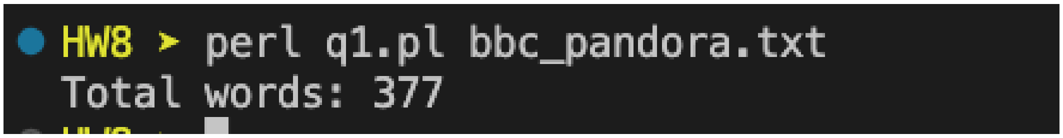
HW8

Q1. In English, names typically begin with an Upper case letter. Other characters may be lower/upper case or include a dash (-), e.g. Al-Ghad. Write a regex and find all the matching words in the article. How many are there?

```
my $file = $ARGV[0];
open (line, $file) or die "Not able to open $file: $! ";
$count = 0;
while (<line>) {
    while (/[A-Z][a-z]+-[A-Z][a-z]+|[A-Z]+[a-z]+/g) {
        $count++;
    }
}
print "Total words: $count\n";
```

Output:

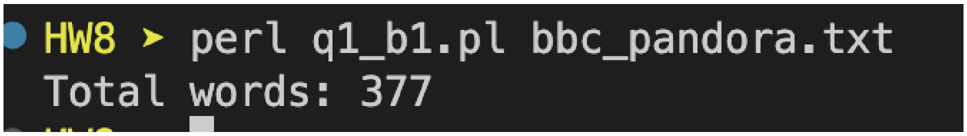


```
HW8 > perl q1.pl bbc_pandora.txt
Total words: 377
```

Q1 Bonus1:

```
my $file = $ARGV[0];
use open qw(:std :utf8);
open (line, $file) or die "Not able to open $file: $! ";
$count = 0;

while (<line>) {
    while (/[A-Z][a-z]+-[A-Z][a-z]+|[A-Z]+[a-z]+/g) {
        $count++;
    }
}
print "Total words: $count\n";
```

output:



```
HW8 ➤ perl q1_b1.pl bbc_pandora.txt
Total words: 377
```

There is no difference in match count probably because there is no word starting with other languages which can not be handled by Unicode.

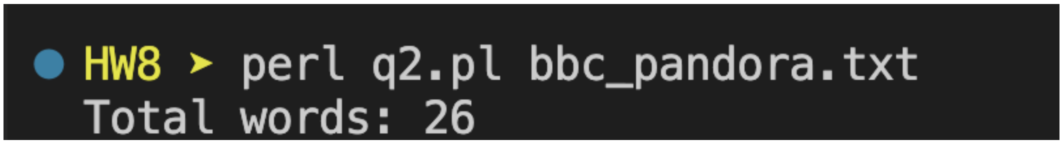
Q1 Bonus 2 Do all name words begin with an Upper case letter? Find one that doesn't.

Almost all of the name words are in uppercase but I found very few like `al-Assad`, `i newspaper`, `#client 13173`

Q2. Abbreviations/acronyms often consist of words, #letters ≥2, containing only Upper case letters, e.g. TV NTV US EPA. Write a regex for this. How many are there?

```
my $file = $ARGV[0];
open (line, $file) or die "Not able to open $file: $! ";
$count = 0;
while (<line>) {
    while (/[A-Z]{2,}/g) {
        $count++;
    }
}
print "Total words: $count\n";
```

Output:

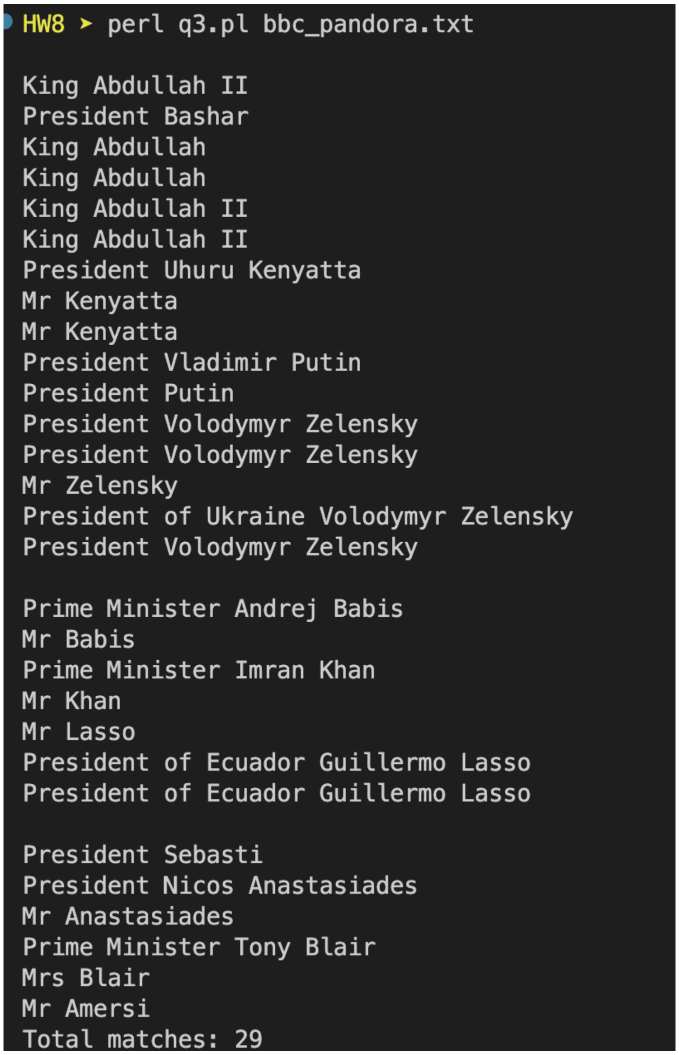


```
HW8 ➤ perl q2.pl bbc_pandora.txt
Total words: 26
```

Q3. Many names are n-grams, for n≥2, a sequence of words each beginning with an Upper case letter, optionally beginning with a title, e.g. Mr/Ms/Mrs/Dr, Prime Minister, President or King/Queen, e.g. Mr Zelensky, President Vladimir Putin or King Abdullah II. Write a regex and find all the matching sequences (#words ≥2) beginning with a title in the article. Print them. How many are there?

```
my $file = $ARGV[0];
open (lines, $file) or die "Not able to open $file: $! ";
my @matches;

while ($line = <lines>) {
    while ($line =~ /((President (\b(?:[A-Z][a-z]*\b\s*)+)))/g) {
       push @matches, $1;
    }
    while (($line =~ /((President ?of (\b(?:[A-Z][a-z]*\b\s*)+)))/g)){
        push @matches, $1;
    }
    while ($line =~ /((Mr|Mrs|Dr)+ [A-Z][a-z]{2,})/g) {
        push @matches, $1;
    }
    while ($line =~ /(Prime Minister [A-Za-z]{2,} [A-za-z]{2,})/g) {
        push @matches, $1;
    }
    while ($line =~ /((King|Queen) (\w+) (I){0,3})/g) {
        push @matches, $1;
    }
}
$count = 0;
foreach $match (@matches) {
    print "$match \n";
    $count++;
}
print "Total matches: $count \n";
```

Output:

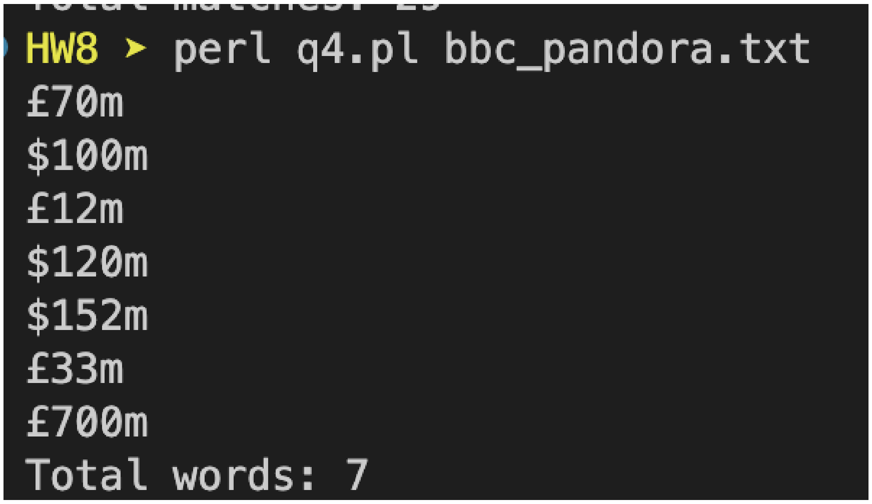


```
HW8 ➤ perl q3.pl bbc_pandora.txt

King Abdullah II
President Bashar
King Abdullah
King Abdullah
King Abdullah II
King Abdullah II
President Uhuru Kenyatta
Mr Kenyatta
Mr Kenyatta
President Vladimir Putin
President Putin
President Volodymyr Zelensky
President Volodymyr Zelensky
Mr Zelensky
President of Ukraine Volodymyr Zelensky
President Volodymyr Zelensky

Prime Minister Andrej Babis
Mr Babis
Prime Minister Imran Khan
Mr Khan
Mr Lasso
President of Ecuador Guillermo Lasso
President of Ecuador Guillermo Lasso

President Sebasti
President Nicos Anastasiades
Mr Anastasiades
Prime Minister Tony Blair
Mrs Blair
Mr Amersi
Total matches: 29
```

Q4 write a regex to find all the monetary values quoted in the article. Note currency symbols, comma separators, and abbreviations such as m for million. Print them. How many are there?

```
my $file = $ARGV[0];
open (line, $file) or die "Not able to open $file: $! ";
$count = 0;
while (<line>) {
    while (/\$[0-9]+m|\£[0-9]+m/g) {
        print "$& \n";
        $count++;
    }
}
print "Total words: $count\n";
```

Output:



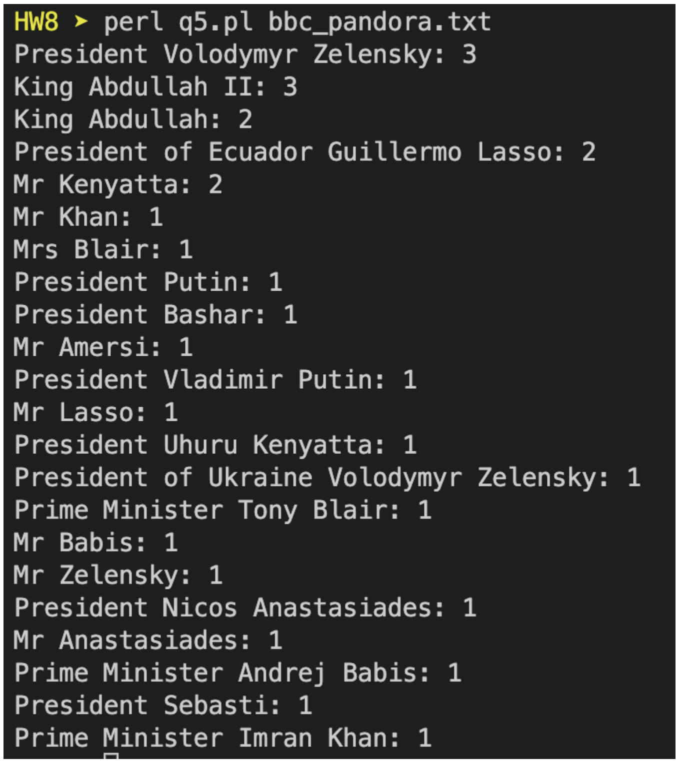
```
Total matches: 23
HW8 ➤ perl q4.pl bbc_pandora.txt
£70m
$100m
£12m
$120m
$152m
£33m
£700m
Total words: 7
```

Q5. Using the Perl hash table described in a previous lecture, re-do question 3 and collect together mentions of names, e.g. King Abdullah occurs multiple times. Then print names and number of occurrences in tabular form, e.g. • Mr Piñera 2

```
my $file = $ARGV[0];
open (lines, $file) or die "Not able to open $file: $! ";
my @matches;

while ($line = <lines>) {
    while ($line =~ /(Prime Minister [A-Za-z]{2,} [A-za-z]{2,})/g) {
        push @matches, $1;
    }
    while ($line =~ /((King (\w+) (I){0,3}))/g) {
        push @matches, $1;
    }
    while ($line =~ /((President (\b(?:[A-Z][a-z]*\b\s*)+)))/g) {
       push @matches, $1;
    }
    while (($line =~ /((President ?of (\b(?:[A-Z][a-z]*\b\s*)+)))/g)){
        push @matches, $1;
    }
    while ($line =~ /((Mr|Mrs|Dr)+ [A-Z][a-z]{2,})/g) {
        push @matches, $1;
    }

}
```

```
my @trim_matches;
foreach $match (@matches) {
    $match =~ s/^\s+//;
    $match =~ s/\s+$//;
    push @trim_matches, $match;
}
my %hash_table;
foreach $match (@trim_matches) {
    $hash_table{$match}++;
}
my @sorted = sort { $hash_table{$b} <=> $hash_table{$a} } keys %hash_table;
foreach $match (@sorted) {
    print "$match: $hash_table{$match} \n";
}
```

Output:



```
HW8 ➤ perl q5.pl bbc_pandora.txt
President Volodymyr Zelensky: 3
King Abdullah II: 3
King Abdullah: 2
President of Ecuador Guillermo Lasso: 2
Mr Kenyatta: 2
Mr Khan: 1
Mrs Blair: 1
President Putin: 1
President Bashar: 1
Mr Amersi: 1
President Vladimir Putin: 1
Mr Lasso: 1
President Uhuru Kenyatta: 1
President of Ukraine Volodymyr Zelensky: 1
Prime Minister Tony Blair: 1
Mr Babis: 1
Mr Zelensky: 1
President Nicos Anastasiades: 1
Mr Anastasiades: 1
Prime Minister Andrej Babis: 1
President Sebasti: 1
Prime Minister Imran Khan: 1
```