Exercise Set 2

Enrico Buratto

Question 1

Elias codes

- **100**: the binary representation for 100 is 1100100; the offset is then 100100 (the binary number without the leading bit). The selector size is 6; in unary code, this is 0000001. The Elias-γ code for 100 is then **0000001100100**.
- **200**: The offset is 1001000, the selector is 7 so 00000001. The Elias-γ code for 200 is then **000000011001000**.
- **400**: The offset is 10010000, the selector is 8 so 000000001. The Elias-γ code for 200 is then **00000000110010000**.

VByte codes

- **100**: the binary representation for 100 is 1100100. This fits 7 bit, so we can put the continuation bit $c = 0$. The VByte code for 100 is then **01100100**.
- **200**: the binary representation for 200 is 11001000. This does not fit 7 bit, so we put the lower 7 bits in the first byte, we put the continuation bit $c = 1$ and we encode the remaining bit in the second byte (with $c = 0$). The VByte code for 200 is then **11001000 00000001**.
- **400**: the binary representation for 400 is 110010000. This does not fit 7 bit, so we put the lower 7 bits in the first byte, we put the continuation bit $c = 1$ and we encode the remaining two bits in the second byte (with $c = 0$). The VByte code for 400 is then **10010000 00000011**.

Question 2

a. In a Boolean retrieval system, stemming never lowers precision - **False**: the retrieved results are more and more general or, at least, the same. The precision is then at most equal, if not lower.
b. In a Boolean retrieval system, stemming never lowers recall - **True**: for the same reason, the recall is at least equal, if not higher.
c. Stemming increases the size of the vocabulary - **False**: more original words can become the same but not vice versa, so the size of the vocabulary is at most equal if not lower.
d. Stemming should be invoked at indexing time, but not while processing a query - **False**: it should be applied also while processing a query in order to match the searched terms with the indexed ones.

Question 3

I would say that **marketing/market** and **university/universe** should not be the same. While abandon/abandonment, absorbency/absorbent and volume/volumes all belong to the same field of interest, `marketing` has a totally different meaning than `markets` (and both are stemmed to `market`); the same applies for `university` and `universe` (that are stemmed to `univers`).