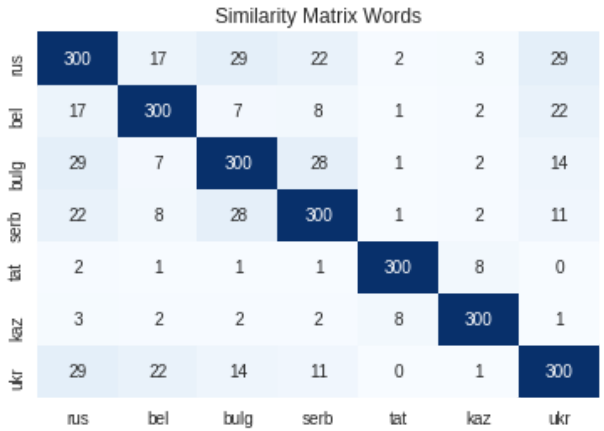

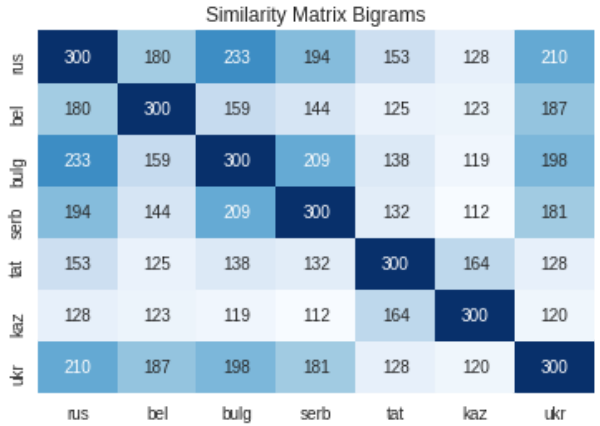

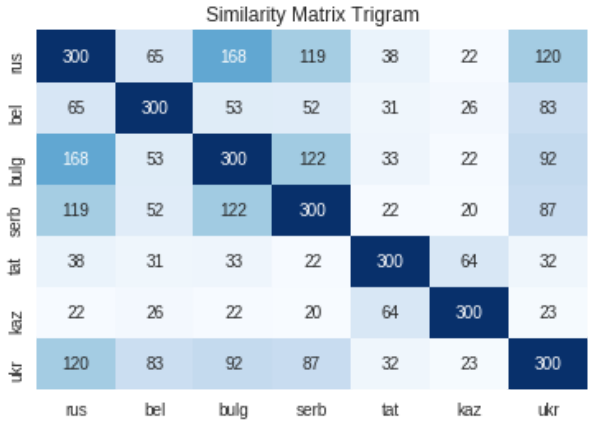

Какие языки оказались очень похожи друг на друга, а какие нет? Иными словами, какие языки проще различить по нграммам, а какие труднее? Интерпретируйте результат.

На основе полученных данных можно сделать ожидаемые выводы, что русский очень схож с украинским, болгарским и сербским языками, а казахский и татарский языки, напротив, ощутимо отличаются от языков славянской группы.

Особенно это сходство заметно при анализе биграмм: у болгарского и русского языков 233 общих высокочастотных биграмма из 300, у украинского с русским - 210, у сербского и болгарского - 209. Это очень высокие показатели, которые указывают на высокую степень схожести языков, что сильно затруднит определение языка по биграммам. Количество общих высокочастотных биграмм у казахского и татарского языков значительно ниже - в диапазоне от 112 - 164, однако это тоже довольно высокие показатели (почти половина от общего количества биграмм). Причем между собой татарский и казахский довольно сильно похожи.

В случае с триграммами картина выглядит несколько лучше, однако языки, демонстрировавшие высокую степень сходства между собой на основе анализа биграм, в данном случае также обладают высокими показателями: на первом месте количество общих триграмм у русского и болгарского языка - 168, также сильно похожи болгарский и сербский - 122 общих триграмма. Точно также казахский и татарский языки на фоне общий картины имеют меньшую степень схожести с другими языками, и при этом демонстрируют высокую степень сходства между собой.

Анализ частотных слов выглядит наиболее оптимальным для определения языка, так как общих высокочастотных слов у этих языков значительно меньше, чем общих биграмм и триграмм: максимальный показатель - 29.
При этом матрица схожести языков еще раз подтверждает выводы, сделанные на основе анализа общих биграмм и триграмм: больше всего похожи между собой русский и болгарский (возможно, это объясняется практически одинаковым алфавитом в этих двух языках), сильное сходство русского, сербского, украинского между собой, и ощутимая дистанция татарского и казахского языков.

Выводы:

- анализ трех матриц показал, что языки, принадлежащие славянской группе, ожидаемо демонстрируют высокую степень схожести;
- на основе сравнения трех матриц между собой можно сделать вывод о том, что менее всего целесообразно определять язык по биграммам, а наиболее оптимально - по отдельным высокочастотным словам.