a. As per the hint, consider a random variable

$$X_{u,v \in S} = \begin{cases} 1 \; if \; h(u) = h(v) \\ 0 \; otherwise \end{cases}$$

Note that $\sum_i n_i^2 = \sum_{u,v \in S} X_{u,v}$ because for any two elements $u \neq v$, we have two random variables $X_{u,v}$ and $X_{v,u}$ and we also have $X_{u,u}$ for any element $u \in S$. That makes for

$$2 * \binom{n}{2} + n = 2 * \frac{n(n-1)}{2} + n = n^2.$$

Thus, $P\left[\sum_{i=0}^{p-1} n_i^2 > 4p\right] \leq \frac{1}{2}$ is the same as $P\left[\sum_{u,v \in S} X_{u,v} > 4p\right] \leq \frac{1}{2}$.

Markov's inequality is

$$P(X \geq a) \leq \frac{E(X)}{a}$$

We now try to get the expectation of the summation.

$$E(\sum_{u,v \in S} X_{u,v}) = E(\sum_{\substack{u,v \in S \; such \\ that \; u \neq v}} X_{u,v}) + E(\sum_{u \in S} X_{u,u})$$

$$= \sum_{\substack{u,v \in S \; such \\ that \; u \neq v}} 1 * P(X_{u,v} = 1) + 0 * P(X_{u,v} = 0) + \sum_{u \in S} 1 * P(X_{u,u} = 1) + 0 * P(X_{u,u} = 0)$$

$$= \sum_{\substack{u,v \in S \; such \\ that \; u \neq v}} 1 * P(X_{u,v} = 1) + \sum_{u \in S} 1 * P(X_{u,u} = 1)$$

$$= 2 * \sum_{i=1}^{\binom{n}{2}} 1 * \frac{1}{p} + \sum_{j=1}^{n} 1 * 1$$

$$= 2 * \frac{n(n-1)}{2} * \frac{1}{p} + n$$

Note that $n = p$

$$= 2p - 1$$

Let's call $\sum_{u,v \in S} X_{u,v}$ X_{new}. If we set the a in Markov's inequality to $4p$, Then

$$P(X_{new} \geq 4p) \leq \frac{2p-1}{4p} < \frac{2p}{4p} = \frac{1}{2}$$

Because $p > 0$.

So

$$P(X_{new} \geq 4p) \leq \frac{1}{2}$$

.

$$P(X_{new} = 4p) \geq 0$$

And

$$P(X_{new} \geq 4p) = P(X_{new} > 4p) + P(X_{new} = 4p)$$

So

$$P(X_{new} > 4p) \leq P(X_{new} \geq 4p)$$

Which finally leads to

$$P(X_{new} > 4p) \leq \frac{1}{2}$$

QED.

b. We first will determine how to get the number of collisions at a table entry i given there are n_i words stored there. A collision is defined as $u, v \in S \textit{ such that } u \neq v \textit{ and } h(u) = h(v)$. If there is one word stored there, there are no collisions; if there are two stored, there is one collision; if there are n_i words stored, there are $\binom{n_i}{2}$ collisions because there's a collision between every pair of words. We now thus define a random variable X that represents the total number of collisions in the hash table:

$$X = \sum_{i=1}^{p} \binom{p_i}{2}$$

Simply:

$$X = \sum_{i=1}^{p} \frac{p_i(p_i - 1)}{2}$$

$$= \frac{1}{2} \sum_{i=1}^{p} p_i(p_i - 1)$$

$$= \frac{1}{2} \sum_{i=1}^{p} p_i^2 - p_i$$

$$= \frac{1}{2}\left(\sum_{i=1}^{p} p_i^2 - \sum_{i=1}^{p} p_i\right)$$

Notice that the first term inside the parenthesis evaluates to the same thing as part a's $\sum_{i=0}^{p-1} n_i^2$ because the empty cells in our now larger hash table don't matter to the summation – only the non-empty cells where the n words reside matter. Taking the expectation of X:

$$E(X) = \frac{1}{2}E\left(\sum_{i=1}^{p} p_i^2 - \sum_{i=1}^{p} p_i\right)$$

$$= \frac{1}{2}\left(E\left(\sum_{i=1}^{p} p_i^2\right) - E\left(\sum_{i=1}^{p} p_i\right)\right)$$

use part a's work:

$$= \frac{1}{2}\left(2 * \frac{n(n-1)}{2} * \frac{1}{p} + n - E\left(\sum_{i=1}^{p} p_i\right)\right)$$

use the fact that $p \geq n^2$, hence the $\leq$ sign:

$$\leq \frac{1}{2}\left(2 * \frac{n(n-1)}{2} * \frac{1}{n^2} + n - E\left(\sum_{i=1}^{p} p_i\right)\right)$$

$$= \frac{1}{2}\left(2 * \frac{n(n-1)}{2} * \frac{1}{n^2} + n - n\right)$$

$$= \frac{1}{2}(n-1) * \frac{1}{n}$$

Note that $n - 1 < n$, so $\frac{n-1}{n} < 1$, so

$$E(X) < \frac{1}{2}$$

The problem wanted us to show that there's at least $\frac{1}{2}$ probability that there's no collisions, which is logically equivalent to there's less than $\frac{1}{2}$ probability that there is a collision. $E(X)$ is the expected number of collisions. QED.

c. It is known from the previous parts that $P\left(\sum_{i=1}^{p} n_i^2 \leq 4p\right) \geq \frac{1}{2}$, and $P(no\ collisions\ in\ h_i) \geq \frac{1}{2}$ for all secondary level hash functions h_i. So, it takes an expected 2 tries in getting a good first level hash function and an expected 2 tries in getting a good second level hash function for each cell that has collisions. Computing $h(u)$ for all $u \in S$ and computing $h_i(v)$ for all $v \in T_i$ is $O(n)$. There cannot be more than n total entries in all the secondary hash tables added together. The time complexity in finding an overall good hash function is thus $O(2n) + O(2n) = O(4n) = \boldsymbol{O(n)}$.